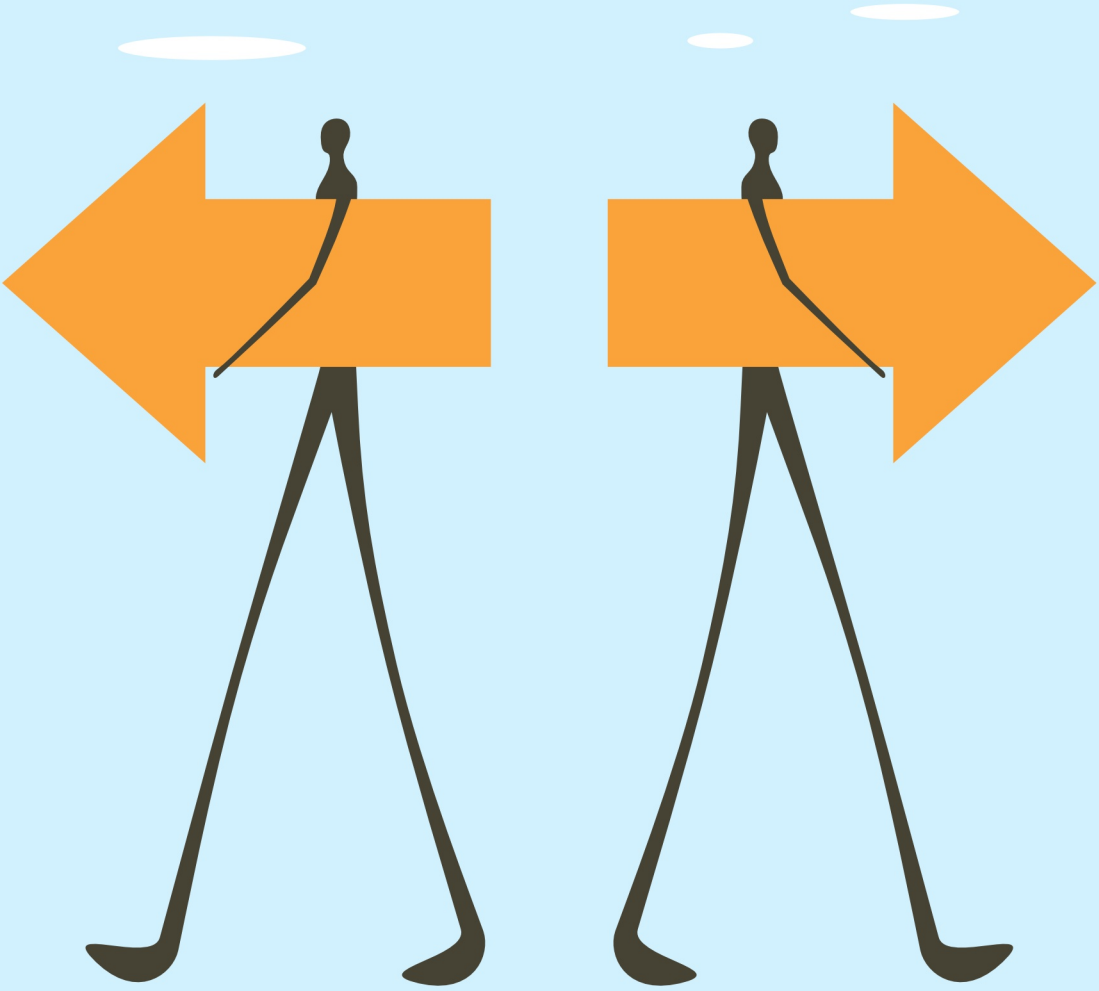
When Data Compression and Statistics Disagree

Two Frequentist Challenges for
the Minimum Description Length Principle



Tim van Erven

When Data Compression and Statistics Disagree

Two Frequentist Challenges for
the Minimum Description Length Principle

When Data Compression and Statistics Disagree

Two Frequentist Challenges for
the Minimum Description Length Principle

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. P. F. van der Heijden,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 23 november 2010
klokke 13.45 uur

door

Tim Adriaan Lambertus van Erven

geboren te Eindhoven

in 1982

Samenstelling van de Promotiecommissie

Promotor: prof. dr. P. D. Grünwald

Overige leden:

prof. dr. A. R. Barron	(Yale University)
dr. P. Harremoës (lic. scient. et exam. art.)	(Niels Brock Copenhagen Business College)
prof. dr. A. W. van der Vaart	(Vrije Universiteit)
prof. dr. P. Stevenhagen	

Printed and bound by Ipskamp Drukkers, Enschede, the Netherlands
ISBN: 978-90-9025673-3

Parts of this thesis are based on the following papers and collaborations.

Chapter 2 is based on:

> *Catching up faster by switching sooner: A predictive approach to adaptive estimation with an application to the AIC-BIC dilemma.* T. van Erven, P. Grünwald, and S. de Rooij. Submitted to the Journal of the Royal Statistical Society, Series B, 2010.
>
> *Catching up faster in Bayesian model selection and model averaging.* T. van Erven, P. D. Grünwald, and S. de Rooij. In: Advances in Neural Information Processing Systems 20 (NIPS 2007), pages 417–424. MIT Press, 2008.

Chapter 3 is based on:

> *Learning the switching rate by discretising Bernoulli sources online.* S. de Rooij and T. van Erven. In: Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS), volume 5 of JMLR: W&CP, pages 432–439, 2009.

Chapter 4 is based on the following technical reports:

> *Switching between hidden Markov models using fixed share.* W. M. Koolen and T. van Erven. Available from http://arxiv.org/abs/1008.4532, 2010.
>
> *Freezing and sleeping: Tracking experts that learn by evolving past posteriors.* W. M. Koolen and T. van Erven. Available from http://arxiv.org/abs/1008.4654, 2010.

Chapter 6 is based on:

> *Rényi divergence.* T. van Erven and P. Harremoës. Manuscript in preparation.

Some of the results in Chapter 6 have already appeared in:

> *Rényi divergence and majorization.* T. van Erven and P. Harremoës. In: IEEE International Symposium on Information Theory (ISIT), pages 1335–1339, 2010.

These investigations were carried out at the Centrum Wiskunde & Informatica (CWI). They were supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886, and by the Thomas Stieltjes Institute for Mathematics. This publication only reflects the author's views.

THOMAS STIELTJES INSTITUTE
FOR MATHEMATICS

Preface

In psychology it is commonly known that by studying pathological cases, one gains more insight into normal functioning. For example, one may think of testing which functions are affected in a patient who has suffered brain damage to the frontal lobe. This does not just help in treating the patient, but also gives important insight into the tasks performed by the frontal lobes of ordinary people, without brain damage.

Analogously, I deliberately seek out pathological cases in statistics, in which two views (one based on data compression, the other on the traditional frequentist perspective) appear to be in conflict. In the first part of the thesis, the pathology is called the catch-up phenomenon and a cure based on switching between models is proposed. In addition, two more chapters are included on similar switching approaches. In the second part of the thesis, deviant behaviour of the so-called minimum description length (MDL) estimator is studied. Although the literature contains a cure, it is based on modifying the MDL estimator, which undermines its data compression interpretation. By refining existing techniques, I improve diagnostics of the undesirable behaviour and show that in certain common cases the MDL estimator is well-behaved even without modification. These cases are characterized using a measure of dissimilarity between probability distributions that was introduced by Alfréd Rényi in the nineteen-sixties. Although Rényi's dissimilarity measure has been around for almost fifty years and frequently appears in mathematical proofs, there exists no overview of its technical properties. The second part of the thesis therefore also includes an overview of the properties of Rényi's dissimilarity measure.

I would like to thank two Peters who have had a strong influence on my work. Firstly, writing this thesis was possible only under the guidance of my advisor, Peter Grünwald. I think we share an appreciation for conceptual issues and an interest in the foundations of statistics. Through his lectures, book and personal advice, Peter's views have shaped my thinking in these matters. Secondly, Peter Harremoës has served as a role model in mathematics. His instant lectures (just add question and stir...) and clarity of thought have been an inspiration.

Apart from "Peter", the names of my fellow PhD students at the Centrum Wiskunde & Informatica (CWI), Steven and Wouter, recur as collaborators and in acknowledgements of my papers. I thank them for many friendly discussions. I would also like to thank my other colleagues at CWI, who have made my time here an enjoyable and stimulating experience. In particular Wojciech Kotłowski's views on the practical importance of theoretical analysis have raised my spirits during the final stages of writing this thesis. Martijn Wallage at the University of Amsterdam prompted me to consider the grue paradox (see Example 1.3 in Chapter 1).

Outside of Amsterdam, I have had the pleasure of visiting Bob Williamson and Mark Reid for two months in Canberra, Australia, and for another week in Cambridge, UK. Although my attention to this thesis has slowed down our joint investigations, I hope we can continue to collaborate and complete our study of geometric properties of loss functions.

Finally, such a long-term project would not have been possible without the support and love of my girlfriend, Klara, and my family and friends. I regret my father has not had the chance to see it undertaken. This thesis is dedicated to his memory.

Amsterdam TIM VAN ERVEN
September, 2010

Contents

List of Figures

Chapter 1

Introduction

> *[T]he object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.*
>
> R. A. Fisher, 1922

It has been recognised at least since Fisher [1922] that statistics and information are closely related. After the theory of information got its proper foundation by the seminal work of Shannon [1948], a series of authors have therefore attempted to base statistics directly on information theory.

In Shannon's setup, data sequences are considered random samples from a known probability distribution, and the amount of information they contain is measured by the expected length of their shortest possible description. This expected description length turns out to be uniquely determined by the distribution of the data.

His approach can be extended to nonrandom data sequences by focusing on descriptions in the form of computer programs, from which the data can be reconstructed by a computer. Although there exist many different programming languages in which computer programs can be expressed, the choice of programming language can only change the shortest possible description length by a constant, as was independently discovered by Solomonoff, Kolmogorov and Chaitin in the nineteen-sixties. This constant does not grow with the length of the data se-

quence, and therefore does not matter for sufficiently long sequences [Li and Vitányi, 2008]. Having thus obtained a measure of the amount of information in nonrandom data sequences, Kolmogorov introduced a method to split the description of the data into a structure component, called the minimal sufficient statistic, and a noise component that is indistinguishable from completely random data. For sufficiently long data sequences, this minimal sufficient statistic captures all patterns in the data that can be described by a computer program [Kolmogorov, 1974, Vitányi, 2005, Cover and Thomas, 1991].

Ironically, however, there is no effective way to compute the minimal sufficient statistic itself, so it cannot be used in practice. A practical variation based on minimizing the description length of the data was therefore proposed by Rissanen [Rissanen, 1978, 1983, 1989, 2007, Grünwald, 2007].[1] Rather than restricting attention to computer programs, this *minimum description length* (MDL) approach relies on a set of probability distributions to determine the language in which the data can be described. The set may be a parametric statistical model, in which case MDL can be used for parameter estimation; or it can be the union of multiple such models, in which case MDL can be used both to select the model (structure) and to estimate its parameters; or the set of distributions may even be nonparametric. This approach was reconnected with random data sequences by findings mainly due to Barron, Rissanen and Yu [Barron and Cover, 1991, Barron, Rissanen, and Yu, 1998], who showed that the MDL estimator satisfies certain statistical properties that Fisher would appreciate. In particular, it is consistent, and automatically prevents overfitting complex models to the data, in the sense that the models fit the data well but lead to poor predictions on unseen data from the same source. This line of work is continued in the present thesis, in which all topics are related to theoretical properties of the MDL estimator.

Overview of the Thesis The remainder of this chapter introduces the MDL estimator and related ideas, which motivate the developments in the rest of the thesis. Although all chapters can be read independently, for a full appreciation it is therefore recommended to read the present chapter first.

[1] Similar methods where suggested earlier by Wallace and Boulton [1968]. See also [Wallace and Freeman, 1987].

The rest of the thesis is split in two parts. In *Chapter 2* of Part I we investigate cases in which standard MDL model selection leads to suboptimal predictions of future data. It is found that this may be explained by the fact that there exist shorter descriptions of the data than the descriptions used by standard MDL. Based on this insight, we modify the standard MDL estimator such that it can use these shorter descriptions and show that this resolves the problem. As a by-product, our investigations shed new light on an old discussion in statistics about whether one should use an AIC-type method or a BIC-type method for model selection. (The details of this debate will be introduced in Chapter 2.)

The shorter descriptions found in Chapter 2 are based on combinations of the models that use a different model for different parts of the data. In *Chapter 3* a new method is introduced that automatically determines the optimal bias towards splitting the data into more parts. In *Chapter 4* we discuss whether the parts should be modelled independently, or as part of the rest of the data. A new method is introduced to deal with the first case, which is appropriate, for example, for certain time series data.

In Part II we also study the quality of predictions based on the MDL estimator, and investigate under which conditions they converge to the best possible predictions. In order to prove a very general convergence result, previous authors have proposed to modify the standard estimator in a way that, contrary to its design philosophy, *increases* the description length of the data (see Section 1.3.4). *Chapter 5* provides a preliminary discussion of whether this modification is really necessary. Examples are provided showing that no general convergence result can be obtained if the modification is simply omitted, but then it is also shown that in certain common settings no modification is necessary. These settings are characterized using a measure of dissimilarity between probability distributions called Rényi divergence [Rényi, 1961]. Although Rényi divergence has been around for almost fifty years and appears in many proofs, there exists no overview of its technical properties. *Chapter 6* remedies this situation by formally proving the basic properties of Rényi divergence.

A more detailed outline of the thesis is provided in Section 1.5, at the end of this chapter.

Overview of Chapter 1 We will proceed to define the MDL estimator and discuss its possible motivations in the next section. Then, in Section 1.2, we will introduce the required information theoretic background on description lengths, before discussing the MDL estimator in the context of parameter estimation in Section 1.3. In Section 1.4 the estimator is extended to model selection, which is its most common area of application. The chapter concludes with an outline of the remainder of the thesis.

1.1 On Minimizing Description Length

Given a countable set of densities $\mathcal{M} = \{p_1, p_2, \ldots\}$, which we will call a *(statistical) model*, and data D, the MDL estimator selects the density that achieves

$$\min_{p \in \mathcal{M}} \left\{ L(p) - \log p(D) \right\}, \tag{1.1}$$

where the logarithm is to base 2. As discussed below, the nonnegative numbers $L(p)$ satisfy Kraft's inequality, $\sum_p 2^{-L(p)} \leq 1$, and are interpreted as the description lengths (or code lengths as they will later be called) of the densities. Note that higher density $p(D)$, which means a better fit on the data, implies that $-\log p(D)$ is smaller. MDL therefore trades off the fit of p on the data with the complexity of p, as measured by $L(p)$. The choice of $L(p)$ and extensions to uncountable models will be discussed in Section 1.3.5.

The minimum description length estimator gets its name from the fact that $L(p) - \log p(D)$ may be regarded as the length of a two-part description of the data, as explained in Section 1.3.2. Here $L(p)$ represents the relevant information in the data, and $-\log p(D)$ represents the noise. MDL's choice for the shortest such description may be motivated in three ways.

The Data Compression Motivation First, some authors, most notably Rissanen [2007], argue that finding the shortest possible description of the data should be taken as the main goal of statistical inference. MDL may then be viewed as an attempt to achieve this goal, subject to the constraint that descriptions are of the form $L(p) - \log p(D)$. We will call this the *data compression* motivation for MDL. Note that it leaves open the possibility that descriptions taking a different form may

be shorter and should therefore be preferred. The data compression motivation is appealing because it incorporates in a very direct way the statistical objective expressed by Fisher of representing the data by fewer quantities that adequately represent the whole: the amount of information in the data is (1.1); then the noise is discarded and the relevant information (the identity of a density from $\mathcal{M}$) is retained. We see that $\mathcal{M}$ determines not only which information is relevant, but also how much information is present in the first place. Ideally, to fully explain the data, the model $\mathcal{M}$ should therefore make the description length (1.1) as small as possible.

The argument for data compression is based on the fact that any regularity in the data may be used to reduce its description length [Grünwald, 2007, Chapter 1]. Minimizing description length, then, is an attempt to capture as much regularity as possible. For example, it is well-known from information theory that any known probabilistic pattern in the data can be used to shorten their description: the less uniform their distribution, the more succinctly the data can be described. Informally, the same phenomenon can also be observed in natural language, in which the number

"one million"

can be described using fewer letters than the number

"five hundred twenty-four thousand, two hundred eighty-eight",

because it has more structure in the decimal system, which underlies natural language. In applying these ideas, one quickly realises that structure or regularity depends on the language used to describe the data. For example, if natural language had been based on the binary system, then the fact that the second of the two numbers above happens to be 2^{19}, would allow it to be described using fewer letters than the first, which becomes "11110100001001000000" in binary. And if a known probabilistic pattern is to be fully exploited to shorten the description of the data, then the description language must depend on their distribution. As a consequence, it is a modelling decision which language to use. In MDL this choice is determined by the model $\mathcal{M}$.

The Frequentist Motivation The data compression motivation should be considered nonstandard, and probably even controversial, because it interprets probabilities (or rather their negative logarithm) as description lengths instead of limiting relative frequencies, which is their classical *frequentist* interpretation [Wasserman, 2005]. In contrast to MDL, the design of frequentist statistical methods is based on the assumption that the data form a random sample from a hypothetical infinite population [Fisher, 1922], and their quality is judged based on long run frequency properties under assumptions on the nature of this population. For example, a frequentist method may be designed to estimate the density of the true distribution of the data under the assumption that this density is differentiable. However, although modern frequentist methods strive to keep the number of assumptions about the population to a minimum [Wasserman, 2006], they do not resolve two concerns raised by adherents of the data compression point of view. The first concern is that even a relatively weak assumption like differentiability of the true density is already quite strong: for example, if an observed datum is the sum of a large number of independent discrete random variables, then even though it may be approximately normally distributed (by the central limit theorem), its density will still be discontinuous [Grünwald, 2007, Example 17.1]. The second concern is that whether the data form a random sample from the proposed population in the first place, may be impossible to verify [Barron et al., 1998] or in some cases does not even make sense. For example, in Chapter 2 Markov models will be used to model the English text in the famous novel "Alice's Adventures in Wonderland" by Lewis Carroll. Should we really imagine this book to be a random sample from a hypothetical infinite set of books written by Lewis Carroll? Or should the population consist of books by any British author? Or perhaps just books in general, including those in Russian? Certainly the patterns found using Markov chains are different for "Alice's Adventures in Wonderland" than they would be for a Russian text.

In spite of these concerns, it seems hard to argue with the position that *if* the frequentist assumptions apply, then long run frequency guarantees are desirable, and one would rightfully be dissatisfied if they could not be given. Several such guarantees for the MDL estimator appear below, as Theorems 1.3, 1.4 and 1.5. For frequentists these may provide a justification of MDL that does not refer to any descrip-

tion lengths. And from a data compression point of view, they provide valuable insight into the data compression properties of MDL.

The Bayesian Motivation or a Motivation for Bayes Finally, there exists yet another approach to statistics, called *Bayesian* inference, which is very popular in, for example, the field of machine learning [Bishop, 2006]. Suppose $\mathcal{M} = \{p_\theta \mid \theta \in \{1, 2, \ldots\}\}$ is a statistical model, indexed by a parameter θ. Then the Bayesian approach assumes that one can always assign so-called *prior probabilities* $\pi(\theta)$ to the possible values of θ. Interpreting $p_\theta(D)$ as the conditional density of data D given the parameter θ, this defines a joint distribution on D and θ with density

$$p(\theta, D) = \pi(\theta) p_\theta(D),$$

on which various types of inference can be based in a coherent way [Bernardo and Smith, 1994]. For example, one may compute the conditional probability that $\theta = 3$ given the observed data D. The Bayesian approach generalises to uncountable and even nonparametric models, and methods for approximate inference exist that make the required computations practical in many cases, including elaborate hierarchical models.

Bayesian inference may have a frequentist interpretation if the prior probabilities are set equal to known relative frequencies of a population, but typically such relative frequencies are not known and the prior is determined either based on *subjective beliefs* or on a *reference analysis* such that its influence on the inference procedure is as small as possible in a certain sense [Bernardo and Smith, 1994]. In these typical cases, Bayesian procedures are controversial, because they do not necessarily give any long run frequency guarantees [Wasserman, 2005].

There is another way to interpret Bayesian inference, however, which is by a formal equivalence with minimum description length methods. In particular, Section 1.3.3 discusses how MDL minimizes the Bayesian probability of error, and in Section 1.4 it is seen how Bayesian model selection with certain objective priors can be regarded as an MDL procedure. Therefore, from a Bayesian perspective one may regard the MDL estimator as a Bayesian estimator, where the choices of L suggested in Section 1.3.5 correspond to objective choices of priors, based on data compression considerations. Alternatively, however, one may also regard MDL as a justification for using these Bayesian meth-

ods, which is meaningful regardless of any prior beliefs. This perspective only applies when Bayes and MDL coincide, and requires that the prior probabilities have good data compression properties. Frequentist results about MDL then transfer to their corresponding Bayesian counterparts. A further comparison between MDL and Bayes is provided by Grünwald [2007, Chapter 17].

1.2 Information Theoretic Preliminaries

The amount of information in an object $x \in \mathcal{X}$ can be measured by the smallest number of symbols from a finite alphabet $\mathcal{A}$ a hypothetical sender, Alice, needs to send to a hypothetical receiver, Bob, to uniquely identify x among all other objects in $\mathcal{X}$. There are two plausible communication models[2], which might be called the *letter model* and the *telegraph model*. In the letter model, Alice sends Bob a letter in which she has written her message using only symbols from $\mathcal{A}$. In the telegraph model, Alice sends her message by first sending the first symbol, then sending the second symbol, and so on, until she comes to the end. To avoid confusion, she has to make clear to Bob when her message ends, for example by sending a special STOP-symbol. We will now formalise these models. Then it will be argued that only the telegraph model is appropriate to measure information. (The restriction to what we call the telegraph model is standard in information theory.) Finally, it will be shown how message lengths in the telegraph model map to probabilities and vice versa.

The Letter Model: Arbitrary Codes We will say that Alice's message *encodes* an object x from among a countable set $\mathcal{X}$ by a corresponding *code word* $s \in \mathcal{A}^* = \bigcup_{\ell=0}^{\infty} \mathcal{A}^\ell$, which is a finite string of elements from $\mathcal{A}$. It is required that code words are unambiguous in the sense that they identify at most one element $x \in \mathcal{X}$. That is, there should exist a *decoding function* $C^{-1}\colon \mathcal{A}^* \to \mathcal{X}$, which maps code words to objects from $\mathcal{X}$ and may be undefined for some code words that are not used.

[2]Here the word "model" is used in its general meaning, and does not refer to the statistical concept of a set of probability distributions, which is used elsewhere in this thesis.

Then, a function C is called a *code*[3] if there exists a decoding function C^{-1} such that C maps any $x \in \mathcal{X}$ to the set $C(x) = \{s \mid C^{-1}(s) = x\}$ of code words that decode to x. The difference between the letter model and the telegraph model lies in which codes they allow. In the letter model, every possible code is allowed.

Example 1.1. Let $\mathcal{X} = \{\text{RED}, \text{GREEN}, \text{BLUE}\}$ and $\mathcal{A} = \{0, 1\}$. Then the following function C is a code: $C(\text{RED}) = \{00\}, C(\text{GREEN}) = \{01\}, C(\text{BLUE}) = \{1\}$. If instead $C(\text{BLUE}) = \{1, 11\}$, then C would also be a code. But if $C(\text{BLUE}) = \{1, 00, 11\}$, then C would not be a code, because the code word 00 could not be unambiguously decoded.

Given a code C, we measure the amount of information in $x \in \mathcal{X}$ by its *code length* $L(x)$, which is defined as the length of the shortest code word for x. That is, $L(x) = \min\{\ell(s) \mid s \in C(x)\}$, where $\ell(s)$ denotes the number of symbols from $\mathcal{A}$ in the code word s. For example, $\ell(01) = 2$. If no code word is associated with x (i.e. $C(x)$ is empty), then we define $L(x) = \infty$.

The Telegraph Model: Prefix-free Codes In the telegraph model Alice and Bob also communicate using a code, but this code has to satisfy an extra requirement: it should always be clear to Bob when Alice is done sending her message. The reason for this, informally, is to disallow messages like:

"$A\ldots$, no wait, I actually meant B!"

when A is also a possible message in itself. In this case Bob cannot decode the message A before knowing that communication has finished. Formally, the restriction imposed by the telegraph model is that codes should be *prefix-free*. That is, there should not exist any two distinct code words s and s' (that are both used) such that s is a prefix of s'. We observe that putting a special STOP-symbol at the end of each code word is one possible (but rather inefficient) way of guaranteeing that a code is prefix-free.

Prefix-free codes have the useful property that the code words for any two prefix-free codes may be *concatenated* to form a new prefix-free

[3] As discussed by Grünwald [2007, p. 80], the definition of a code differs between various standard texts on information theory. The present definition essentially follows [Li and Vitányi, 2008].

code. That is, if $C_{\mathcal{X}}$ and $C_{\mathcal{Y}}$ code for objects x and y from $\mathcal{X}$ and $\mathcal{Y}$, respectively, with code lengths $L_{\mathcal{X}}(x)$ and $L_{\mathcal{Y}}(y)$, then $C_{\mathcal{X}\times\mathcal{Y}}(x,y) = \{s_x s_y \mid s_x \in C_{\mathcal{X}}(x), s_y \in C_{\mathcal{Y}}(y)\}$ is a prefix-free code for objects from $\mathcal{X} \times \mathcal{Y}$, with code lengths

$$L_{\mathcal{X}\times\mathcal{Y}}(x,y) = L_{\mathcal{X}}(x) + L_{\mathcal{Y}}(y).$$

For example, if $\mathcal{X} = \{\text{RED}, \text{GREEN}, \text{BLUE}\}$ and 11 and 011 are codewords for RED and GREEN, respectively, under a prefix-free code C, then by concatenating C with itself we can encode the sequence RED, GREEN by 11011.

Restriction to Prefix-free Codes At first sight, both the telegraph model and the letter model may seem reasonable ways of measuring the information in an object. However, it turns out that only the telegraph model can ensure that information is *sub-additive*, in the sense that the information in objects x and y separately is never less than the information in (x,y) together. In other words, it should not be possible to transmit x and y using fewer symbols using two messages, than it takes to transmit them in a single message. Therefore only the telegraph model is appropriate to measure information.

To make this argument precise, suppose $C_{\mathcal{X}}$ and $C_{\mathcal{Y}}$ encode objects x and y from countable sets $\mathcal{X}$ and $\mathcal{Y}$, respectively, with code lengths $L_{\mathcal{X}}$ and $L_{\mathcal{Y}}$. Then if $L_{\mathcal{X}}(x)$ and $L_{\mathcal{Y}}(y)$ are reasonable measures of the amount of information in x and y, there should exist a code $C_{\mathcal{X}\times\mathcal{Y}}$ to encode objects $(x,y) \in \mathcal{X} \times \mathcal{Y}$ such that

$$L_{\mathcal{X}\times\mathcal{Y}}(x,y) \leq L_{\mathcal{X}}(x) + L_{\mathcal{Y}}(y) \qquad \text{(sub-additivity)} \tag{1.2}$$

for all x and y.

For the telegraph model it is easy to construct a code $C_{\mathcal{X}\times\mathcal{Y}}$ that satisfies (1.2) with equality, simply by concatenating $C_{\mathcal{X}}$ and $C_{\mathcal{Y}}$ as described above. The letter model, however, does not satisfy sub-additivity, as shown by the following counterexample.

Example 1.2. Observe that sub-additivity implies that, for any code $C_{\mathcal{X}}$ and any integer n, there should exist a code $C_{\mathcal{X}^n}$ such that

$$L_{\mathcal{X}^n}(x^n) \leq \sum_{i=1}^{n} L_{\mathcal{X}}(x_i), \quad \text{for all } x^n = x_1, \ldots, x_n \in \mathcal{X}^n. \tag{1.3}$$

m	5	6	7	8	9	10
$\binom{5}{m-5}2^5$	32	160	320	320	160	32

Table 1.1: Counts from Example 1.2

Consider now $C_{\mathcal{X}}(a) = \{1\}, C_{\mathcal{X}}(b) = \{11\}, C_{\mathcal{X}}(c) = \{0\}, C_{\mathcal{X}}(d) = \{00\}$ for $\mathcal{X} = \{a, b, c, d\}$ and $\mathcal{A} = \{0, 1\}$. For this code there are $\binom{n}{m-n}2^n$ choices of x^n such that $\sum_{i=1}^n L_{\mathcal{X}}(x_i) = m$. Table 1.1 tabulates this for $n = 5$. We see there are 192 sequences x^5 such that $\sum_{i=1}^5 L_{\mathcal{X}}(x_i) \leq 6$. However, there are only $2^7 - 1 = 127$ code words of length at most 6. Therefore, there does not exist a code $C_{\mathcal{X}^n}$ that achieves (1.3) and we conclude that the letter model does not satisfy sub-additivity.

In light of the above, we adopt the telegraph model, which corresponds to restricting ourselves to prefix-free codes. (This restriction is standard in information theory [Cover and Thomas, 1991][4].) In the sequel, when we say code, we will actually mean prefix-free code.

Code Lengths are Probabilities There is a fundamental limit to how many objects from $\mathcal{X}$ can be assigned short code lengths. This limit is expressed by Kraft's inequality [Cover and Thomas, 1991]:

Theorem 1.1 (Kraft's Inequality)**.** *Let $a = |\mathcal{A}|$ denote the number of available coding symbols. Then the code lengths of any prefix-free code satisfy*

$$\sum_{x \in \mathcal{X}} a^{-L(x)} \leq 1. \tag{1.4}$$

Conversely, for any function $L \colon \mathcal{X} \to \mathbb{N}$ that satisfies (1.4) there exists a prefix-free code with code lengths equal to L.

Kraft's inequality suggests a correspondence between codes and probability distributions: consider a nonnegative function p on $\mathcal{X}$ such that

$$\sum_{x \in \mathcal{X}} p(x) \leq 1. \tag{1.5}$$

Such functions are called *probability mass functions*. If (1.5) holds with equality, then p defines an ordinary probability distribution on $\mathcal{X}$. We

[4] Although it is usually motivated differently, using an argument based on unique decodability of the concatenation of a code with itself.

will call such ordinary distributions *complete*. Alternatively, if the inequality in (1.5) is strict, then p still defines a measure on $\mathcal{X}$, which we will call an *incomplete* distribution. One may think of incomplete distributions as complete distributions with some probability mass on an extra object outside of $\mathcal{X}$. They are commonly used in information theory, for example because they simplify axiomatic characterizations of measures of entropy and information [Rényi, 1961].

The correspondence suggested by Kraft's inequality can now be formulated as follows: for any code with code lengths $L(x)$, $p(x) = a^{-L(x)}$ is a probability mass function that defines a (possibly incomplete) probability distribution. And vice versa, for any (possibly incomplete) distribution with probability mass function p, there exists a code with code lengths $L(X) = \lceil -\log_a p(x) \rceil$. Here $\lceil z \rceil$ denotes rounding up z to the nearest integer. Rounding up $-\log p(x)$ is necessary because code lengths are restricted to be integers by definition. In statistical or data compression applications, however, $-\log p(x)$ will typically be so large that the effect of rounding is negligible and can easily be ignored. For example, if $x = x_1, \ldots, x_n$ is a sample of size n, then $-\log p(x)$ will typically be linear in n. Adopting therefore this minor idealisation, we find that code lengths and probabilities become formally *equivalent*:

Definition 1.1 (Idealised Code Lengths)**.** A function $L\colon \mathcal{X} \to \mathbb{R}$ is called an *(idealised) code length* function if

$$L(x) = -\log_a p(x) \qquad \text{for all } x \in \mathcal{X} \tag{1.6}$$

for some (possibly incomplete) probability mass function p on $\mathcal{X}$, where $a = |\mathcal{A}|$ denotes the number of available coding symbols.

Apart from a constant multiplication factor $1/\log(a)$, this definition is independent of the choice of $\mathcal{A}$, which makes choosing the base of the logarithm a matter of convenience. By default we will take $a = 2$, such that code length is measured in *bits*. But sometimes it will be convenient to use $a = e$ to get the natural logarithm, for which code length is measured in *nats*. Note that the larger $p(x)$, the smaller $L(x)$, and that $L(x)$ is never negative.

The correspondence between code lengths and probabilities from Definition 1.1 is not just of a syntactic nature. For any distribution, the corresponding code length function uniquely achieves the minimum code length in expectation, which is called the *entropy* of the distribution [Cover and Thomas, 1991, Theorems 5.3.1 and 5.4.3]:

Theorem 1.2. *If X is distributed according to P, then for any (idealised) code length function L*

$$\mathbf{E}[L(X)] \geq \mathbf{E}[-\log P(X)],$$

with equality if and only if $L(X) = -\log P(X)$.

A similar result holds in probability [Cover and Thomas, 1991, Theorem 5.11.1]. The upshot of this section, therefore, is that (idealised) code lengths and probabilities are equivalent in a strong sense, and can be identified. Based on this reasoning, in future chapters we often interpret the negative logarithm of probabilities as code lengths. Our results, however, do not rely on this interpretation.

1.3 MDL Parameter Estimation

With the information theoretic preliminaries out of the way, let us move on to fill in some details that were left out when the minimum description length estimator was introduced in Section 1.1. We then present some of its frequentist properties and finally the choice of code lengths for the densities in the model will be discussed.

1.3.1 MDL Estimator

Let $\mathcal{X}^n$ denote the direct product of n copies of a sample space $\mathcal{X}$, and let $\mathcal{M} = \{p_1, p_2, \ldots\}$ be a countable statistical model, where each $p \in \mathcal{M}$ is a density on $\mathcal{X}^n$ with respect to a common σ-finite dominating measure μ. We use the corresponding upper-case letter (e.g. P) to refer to the distribution corresponding to a density (e.g. p). An *estimator* is a measurable function $\hat{p}\colon \mathcal{X}^n \to \mathcal{M}$ that maps any data $x^n \in \mathcal{X}^n$ to an element $\hat{p}(x^n)$ of the model $\mathcal{M}$. For example, the *maximum likelihood estimator* is defined as

$$\hat{p}(x^n) = \arg\max_{p \in \mathcal{M}} p(x^n),$$

whenever the maximum $\max_{p\in\mathcal{M}} p(x^n)$ is uniquely achieved.

Let $L\colon \mathcal{M} \to \mathbb{R}$ be an (idealised) code length function. Then the *minimum description length estimator* with density code lengths L is defined as

$$\ddot{p}(x^n) = \arg\min_{p \in \mathcal{M}} \Big\{ L(p) - \log p(x^n) \Big\}.$$

If there are multiple p achieving the minimum, then the one with smallest code length $L(p)$ is selected. Any further ties are resolved arbitrarily, for example by selecting p with smallest index in $\mathcal{M}$. Note that, if $\mathcal{M}$ is finite, then the maximum likelihood estimator is a special case of the MDL estimator, with density code lengths $L(p)$ that are the same for all $p \in \mathcal{M}$.

1.3.2 Coding Interpretation

The main interpretation of the MDL estimator is as a minimizer of the length of a two-part description of the data.

Countable Sample Space To give the precise interpretation, suppose first that $\mathcal{X}$ is countable (i.e. the data are discrete) and that each $p \in \mathcal{M}$ is a probability mass function on $\mathcal{X}^n$. Then, for data $x^n \in \mathcal{X}^n$, we may interpret $L_p(x^n) = -\log p(x^n)$ as the (idealised) code length of x^n under the code corresponding to p. Consequently, the data can be described in two parts: first encode p using $L(p)$ bits and then encode x^n using $L_p(x^n)$ bits. For any $p \in \mathcal{M}$, this gives a total description length of

$$L(p) + L_p(x^n) \tag{1.7}$$

bits. Among such descriptions of the data, the minimum description length estimator selects the shortest.

Clearly, neither the model $\mathcal{M}$ nor the choice of density code lengths L is allowed to depend on x^n. To allow otherwise would present the receiver of a message encoding x^n with a *Catch-22* problem: in order to decode the message, he would have to know $\mathcal{M}$ and L, but in order to know both $\mathcal{M}$ and L he would first have to decode the message.

Also note that the MDL estimator does not depend on the actual choice of code words, but only on their lengths. For idealised code lengths these lengths only depend on the alphabet $\mathcal{A}$ through a constant multiplication factor, which does not affect the estimator. Thus the choice of alphabet does not matter, as it should not.

Uncountable Sample Space As there are only a countable number of possible code words, the previous coding interpretation does not directly apply when $\mathcal{X}$ is uncountable, since there are not enough code

words to encode more than a vanishingly small fraction of an uncountable set. Nevertheless, one may regard this as the limiting case of recording the data to increasingly high precision.

Suppose for concreteness that $\mathcal{X} = \mathbb{R}$ (the reasoning generalises to higher dimensions as well) and that densities are with respect to the standard Lebesgue measure μ. Let $[x^n]_d$ denote $x^n \in \mathcal{X}^n$ with each outcome x_i recorded to d decimal places. For given precision d, the MDL estimator prefers $p \in \mathcal{M}$ over $q \in \mathcal{M}$ if

$$\log \frac{Q([x^n]_d)}{P([x^n]_d)} < L(q) - L(p),$$

where $Q([x^n]_d)$ or $P([x^n]_d)$ denotes the probability of the set of data sequences that agree with x^n up to d decimal places. As $p(x^n) = \lim_{d\to\infty} P([x^n]_d)/\mu([x^n]_d)$ almost everywhere, the limiting case as the precision goes to infinity, is

$$\log \frac{q(x^n)}{p(x^n)} < L(q) - L(p)$$

for almost every x^n, which matches the definition of the MDL estimator for uncountable $\mathcal{X}$. Consequently, taking $\mathcal{X}$ to be uncountable corresponds to recording the data to infinite precision.

Remark 1.1. One may regard the supposition that data are recorded to infinite precision as an unrealistic idealisation. Reassuringly, however, Barron [1985] shows that the MDL estimator is well-behaved even if the precision d is taken into account and is allowed to depend on the sample size n. See also the comments by Barron and Cover [1991]. We now leave such issues, as they are outside the scope of this thesis.

1.3.3 Bayesian Interpretation

A secondary interpretation of the MDL estimator can be given from a Bayesian perspective. Let $\mathcal{M} = \{p_1, p_2, \ldots\}$ be a model with a countable number of elements. Each element $p \in \mathcal{M}$ is a density on $\mathcal{X}^n$ with respect to a common σ-finite dominating measure μ. Let π be a prior probability mass function on $\mathcal{M}$ and let $\hat{p}\colon \mathcal{X}^n \to \mathcal{M}$ be an estimator. For any measurable event $A \subseteq \mathcal{X}^n$, let $\mathbf{1}_A$ denote its indicator function, which is 1 on A and 0 otherwise. Then the Bayesian probability of

misidentifying the true density $p \in \mathcal{M}$, drawn randomly according to π, is

$$\sum_p \pi(p) P(\hat{p} \neq p) = \int \sum_p \pi(p) p(x^n) \, \mathbf{1}_{\{\hat{p}(x^n) \neq p\}} \, \mathrm{d}\mu.$$

Consequently, the Bayes estimator, which by definition minimizes this misidentification probability, has to maximize

$$\pi(p) p(x^n) \propto \pi(p \mid x^n) \tag{1.8}$$

almost everywhere, where $\pi(p \mid x^n) = \pi(p) p(x^n) / \sum_p p(x^n) \pi(p)$ denotes the *Bayesian posterior probability* of p given x^n and the $\propto$-relation expresses that two quantities are equal up to a constant multiplication factor. As maximizing (1.8) is equivalent to minimizing

$$-\log \pi(p) - \log p(x^n), \tag{1.9}$$

it follows that the estimator that minimizes the Bayesian misidentification probability, is equal to the MDL estimator with density code lengths $L(p) = -\log \pi(p)$. Based on this correspondence, it is common in the literature to define the density code lengths by specifying a distribution π. Although this distribution π is usually not based on any Bayesian considerations, it is convenient to refer to it as a *prior* nonetheless. In the remainder we will adopt this convention.

MDL is Not Bayes The previous discussion might seem to suggest that MDL is really just Bayes in disguise. However, as will be seen when we come to the selection of π, the coding interpretation leads to choices of priors that cannot usually be reconciled with the belief that a true density is drawn according to such a prior. In particular the optimal MDL priors will often depend on the sample size, and, when model selection is introduced in Section 1.4, it will be seen how MDL leads to procedures that in some cases are even formally non-Bayesian. This section, then, should not be taken as an attempt to justify MDL by giving it a Bayesian interpretation. On the contrary, its point is to show that Bayesian methods (with certain priors) may be justified by reinterpreting them from a coding perspective. Indeed, Grünwald [2007, p. 543] shows that the priors that make Bayesian inference behave badly in an (in)famous example by Diaconis and Freedman [1986], are not acceptable according to the criteria for density code lengths formulated in Section 1.3.5, because they do not compress the data.

1.3.4 Frequentist Properties

The following theorems show that MDL automatically avoids overfitting, regardless of the size or complexity of the model $\mathcal{M}$. This stands in contrast with the behaviour of the maximum likelihood estimator, which needs to be modified by adding appropriate penalizations to complex densities if the model is sufficiently rich.

Let $\mathcal{M} = \{p_1, p_2, \ldots\}$ be a set of densities on $\mathcal{X}$. The densities are extended to multiple outcomes $x^n \in \mathcal{X}^n$ by taking products: $p(x^n) = \prod_{i=1}^n p(x_i)$. Let π be a (possibly incomplete) probability mass function on $\mathcal{M}$, and let $\ddot{p}$ denote the corresponding MDL estimator with density code lengths $L(p) = -\log \pi(p)$. Recall that in this context we refer to π as a prior, even though it need not be based on any Bayesian considerations.

1.3.4.1 Consistency

The following result by Barron and Cover [1991] shows that MDL is consistent if the outcomes are independent and identically distributed (i.i.d.), and the model contains the true density:

Theorem 1.3 (Consistency)**.** *Suppose $X_1, \ldots, X_n$ are drawn independently according to a density $q \in \mathcal{M}$ with finite code length (i.e. $L(q) < \infty$), and the density code lengths do not depend on n. Then*

$$\ddot{p} = q$$

for all sufficient large n, with probability one.

MDL consistency extends to non-i.i.d. settings as long as the distributions in the model are asymptotically sufficiently distinguishable in a suitable sense [Grünwald, 2007, Theorem 5.1]. It is crucial for the consistency of MDL that it takes the density code lengths into account. This is illustrated by considering the way it resolves the *grue paradox* [Goodman, 1955].

Example 1.3 (The Grue Paradox)**.** Let $x_1, \ldots, x_n$ be a sequence of observations of the colour of emeralds, which are assumed to be either green or blue. Let an emerald be *grue* if it is green and observed before the t-th observation is made, or blue and observed after the t-th observation. Likewise, call an emerald *bleen* if it is blue and observed before the t-th

observation is made, or green and observed after the t-th observation. The original paradox casts doubt on whether there is any objective basis, based on observing that $x_1, \ldots, x_n$ are all green[5], to predict that all emeralds are in fact green. As Goodman observes, if t is larger than n, then based on these observations we might equally well predict that all emeralds are grue. Any objection to the extent that green is more plausible than grue, because grue and bleen are defined in terms of green and blue, can be rebutted by noting that blue and green might equally well have been defined in terms of grue and bleen. As formulated by Goodman, there is no escape from the grue paradox. But, if we allow an infinitely continuing series of observations, such that n eventually becomes arbitrarily large, then there does exist an answer, and it is provided by MDL.

To preclude the trivial answer that grue is ruled out as soon as $n > t$, we consider the model $\mathcal{M} = \{p_t \mid t = 1, \ldots, \infty\}$, where p_t assigns probability one to all emeralds being grue, with grue defined relative to t. This ensures that for any n, there exists $t > n$. Formally, let p_t be a point-mass on the infinite sequence of observations that are green up till outcome x_t and blue afterwards, such that $p_t(x^n) = 0$ if $t < n$ and $p_t(x^n) = 1$ otherwise. Note that p_∞ corresponds to the truth that all emeralds are green. Now let $L(p_t)$ be arbitrary density code lengths, which are finite for all t, including $t = \infty$. Then the MDL estimator selects

$$\ddot{p} = \arg\min_{\{p_t:\, t \geq n\}} L(p_t).$$

That is, it selects the simplest density consistent with the observations, where simplicity is measured by $L(p_t)$. Let $\mathcal{S} \subseteq \mathcal{M} \setminus \{p_\infty\}$ denote the set of densities that are at least as simple as the true density, except for the truth itself. Then $L(p_t) \leq L(p_\infty)$ for all $p_t \in \mathcal{S}$ and by Kraft's inequality (1.4) the set $\mathcal{S}$ must be finite. As a consequence $t_{\mathcal{S}} = \max\{t \mid p_t \in \mathcal{S}\}$ is also finite, and for all $n > t_{\mathcal{S}}$ MDL will correctly predict that all emeralds are green. We see that all densities that are simpler than the truth are eventually ruled out as n grows. The simplest remaining density is then the correct one. The reason that MDL is consistent in general is similar: the density code lengths essentially

[5]This observation should come as no surprise, since, according to Wikipedia [Wikipedia entry on *emerald*, 2010], the word *emerald* derives from the Semitic word *izmargad*, which has *green* as its alternative meaning.

restrict the model to a finite set that includes the truth, from which the data then determine the true density. This is most clearly expressed by the proof of Theorem 5.1 in [Grünwald, 2007].

If in fact all emeralds turn out to be grue (for some arbitrary t), then by the same reasoning we see that MDL would also figure this out. This holds regardless of the choice of density code lengths, as long as we make *some* choice. By contrast, the maximum likelihood estimator does *not* resolve the paradox, because it does not provide any way to choose between the densities that are consistent with the data. There is, however, one limitation to MDL's resolution of the grue paradox, which is that for no given n one can be certain that the truth has already been discovered. In the words of Barron and Cover [1991]: "You know, but you do not know you know."

1.3.4.2 Rates of Convergence

Theorem 1.3 shows that MDL will eventually, possibly for very large n, identify the true density. This raises the question of how well MDL approximates the truth for any finite n. Theorem 1.4 below gives an answer. It measures the quality of the MDL approximation in terms of Rényi divergence, under a condition on the tails of the prior.

For any densities p and q on $\mathcal{X}$, let

$$D_\alpha(p\|q) = \tfrac{1}{\alpha-1} \log \int p^\alpha q^{1-\alpha} \, \mathrm{d}\mu$$

denote the *Rényi divergence* (of order α) of p from q. For continuity in α, Rényi divergence of order $\alpha = 1$ is defined equal to the *Kullback-Leibler divergence*

$$D(p\|q) = \mathbf{E}_P \log \frac{p(X)}{q(X)}.$$

Chapter 6 gives an overview of the properties of Rényi divergence. We note already that convergence in $D_{1/2}$ implies convergence in the better known *squared Hellinger distance* $\mathrm{Hel}^2(p,q) = \int (\sqrt{p} - \sqrt{q})^2 \, \mathrm{d}\mu$, because

$$D_{1/2}(p\|q) \geq \mathrm{Hel}^2(p,q).$$

In addition, Rényi divergence is nondecreasing in its order α and for $\alpha = 2$ it is smaller than the χ^2-distance [Gibbs and Su, 2002].

For $\lambda \geq 1$, let $\ddot{p}_\lambda$ denote the *λ-MDL estimator*, defined by

$$\ddot{p}_\lambda(x^n) = \arg\min_{p \in \mathcal{M}} \lambda L(p) - \log p(x^n).$$

For $\lambda = 1$, this is just the ordinary MDL estimator. As will be explained in Chapter 5, other values of λ may be interpreted as applying the ordinary MDL estimator with a prior $w(p) \propto \pi(p)^\lambda$ that satisfies the *light-tails* condition of Barron and Cover [1991]:

$$\sum_{p \in \mathcal{M}} w(p)^{1/\lambda} < \infty.$$

The following result, which is essentially Theorem 15.3 of Grünwald [2007], shows that if the true density can be approximated well by a sufficiently simple element of $\mathcal{M}$, then the density selected by λ-MDL converges to the true density in Rényi divergence.

Theorem 1.4 (Convergence)**.** *Suppose $X^n = X_1, \ldots, X_n$ are distributed i.i.d. according to a density q on $\mathcal{X}$, which need not be a member of $\mathcal{M}$. Let $\hat{p}\colon \mathcal{X}^n \to \mathcal{M}$ be* any *estimator and abbreviate $\hat{p} = \hat{p}(X^n)$. Then for any $\lambda > 1$ and $\varepsilon > 0$*

$$D_\alpha(q \| \hat{p}) \leq \frac{\lambda L(\hat{p}) - \log \hat{p}(X^n) + \log q(X^n)}{n} + \lambda\varepsilon \tag{1.10}$$

with probability at least $1 - e^{-n\varepsilon}$, where $\alpha = 1 - 1/\lambda$. Moreover

$$\mathbf{E}_{X^n} D_\alpha(q \| \hat{p}) \leq \mathbf{E}_{X^n}\left[\frac{\lambda L(\hat{p}) - \log \hat{p}(X^n) + \log q(X^n)}{n}\right]. \tag{1.11}$$

Proof. Let $f(p, x^n) = \Big(p(x^n)/q(x^n)\Big)^{1/\lambda}$ for $p \in \mathcal{M}, x^n \in \mathcal{X}^n$, and for the remainder of this proof adopt the convention that $0/0 = 1$. Then

$$\begin{aligned} 1 \geq \sum_p \pi(p) = \sum_p \pi(p) \frac{\mathbf{E}_{X^n} f(p, X^n)}{\mathbf{E}_{Y^n} f(p, Y^n)} &= \mathbf{E}_{X^n} \sum_p \frac{\pi(p) f(p, X^n)}{\mathbf{E}_{Y^n} f(p, Y^n)} \\ &\geq \mathbf{E}_{X^n} \frac{\pi(\hat{p}) f(\hat{p}, X^n)}{\mathbf{E}_{Y^n} f(\hat{p}, Y^n)} = \mathbf{E}_{X^n} Z(X^n), \end{aligned}$$

where we have introduced the abbreviation

$$Z(X^n) = \frac{\pi(\hat{p}) f(\hat{p}, X^n)}{\mathbf{E}_{Y^n} f(\hat{p}, Y^n)}.$$

As additivity of Rényi divergence (see Chapter 6) implies that

$$\lambda \log Z(X^n) = nD_\alpha(q\|\hat{p}) - \lambda L(\hat{p}) + \log \frac{\hat{p}(X^n)}{q(X^n)},$$

(1.10) follows by rewriting the following application of Markov's inequality:

$$Q\left(Z(X^n) \geq e^{n\epsilon}\right) \leq e^{-n\epsilon}\, \mathbf{E}_{X^n}\, Z(X^n) \leq e^{-n\epsilon}$$

and (1.11) is obtained from

$$\mathbf{E}_{X^n} \log Z(X^n) \leq \log \mathbf{E}_{X^n}\, Z(X^n) \leq 0,$$

which uses Jensen's inequality. □

The bounds of the theorem are optimized by letting $\hat{p}$ be the λ-MDL estimator $\ddot{p}_\lambda$. We see that *the more this estimator compresses the data* (i.e., the smaller $\lambda L(\ddot{p}_\lambda) - \log \ddot{p}_\lambda(x^n)$), *the better it learns*. In particular, the right-hand side of (1.11) goes to zero if the true density q can be approximated well by a sufficiently simple density in $\mathcal{M}$. This is illustrated by the following corollary, which shows that the λ-MDL estimator converges to q at a rate that trades off the complexity $L(p)$ of an approximation $p \in \mathcal{M}$ with the quality of that approximation, measured in terms of the Kullback-Leibler divergence $D(q\|p)$.

Corollary 1.1. *Let $\ddot{p}_\lambda$ be the λ-MDL estimator for $\lambda > 1$, and suppose $X^n = X_1, \ldots, X_n$ are i.i.d. according to a density q on $\mathcal{X}$. Then for $\alpha = 1 - 1/\lambda$*

$$\mathbf{E}\, D_\alpha(q\|\ddot{p}_\lambda) \leq \min_{p \in \mathcal{M}} \left\{ \frac{\lambda L(p)}{n} + D(q\|p) \right\}. \tag{1.12}$$

Consequently, if $q \in \mathcal{M}$ then

$$\mathbf{E}\, D_\alpha(q\|\ddot{p}_\lambda) \leq \frac{\lambda L(q)}{n}. \tag{1.13}$$

Note that, for $\lambda \geq 2$, the theorem and its corollary still hold if Rényi divergence is replaced by the squared Hellinger distance. Unfortunately, they become vacuous as $\lambda \downarrow 1$, corresponding to the ordinary MDL estimator. Thus, MDL estimators based on a prior with "light

tails" converge to the true density, but unfortunately we cannot establish the same result for *arbitrary* MDL estimators. We postpone further discussion of this issue to Chapter 5, where it is the main topic.

The second step of the corollary, (1.13), is really a significant weakening compared to (1.12), because it restricts attention to $q \in \mathcal{M}$. By contrast, (1.12) also applies to q that can only be approximated by elements of $\mathcal{M}$. Although we may consider such q to be infinitely complex: $L(q) = \infty$, they can still be learned as long as $\mathcal{M}$ contains an approximating sequence $p_1, p_2, \ldots$ such that both $D(q\|p_n) \to 0$ and $L(p_n)/n \to 0$. Such approximations underlie applications of MDL in nonparametric settings (see Section 1.4.4).

If $q \in \mathcal{M}$, but $L(q)$ is still so large (relative to the sample size) that (1.13) is vacuous, then for all practical purposes we are in the same case as above, and if a simpler approximation to q exists, it will lead to better predictions. As will be discussed next, this provides a formal justification for *Occam's razor*.

Occam's Razor By definition the MDL estimator trades off goodness-of-fit on the data against complexity of the densities. This can be interpreted as a formalisation of Occam's razor: the heuristic commonly applied in science, which suggests to prefer simple explanations over more complex ones. Occam's razor has sometimes been criticised on the grounds that it represents a naive belief that simple explanations are more likely to be true than complex ones [Domingos, 1999]. Equation 1.12 in Corollary 1.1, however, presents a different motivation for Occam's razor. It shows that simple approximations to the truth lead to better convergence rates and therefore make better predictions of future data, even if the truth is very complex. On the other hand, Equation 1.13, which directly relates convergence to the complexity of the truth, becomes vacuous if the truth is too complex to learn at the current sample size. In conclusion: if the truth is very complex, it is preferable to learn a simple approximation, because this will lead to better predictions on future data. As more data become available, increasingly complex (approximations of the) truth can be considered.

Remark 1.2 (Related Work). Up to a constant multiplicative factor, Theorem 1.4 can also be obtained as a special case of Theorem 2.1 by Zhang [2006], which is based on a convex duality used in PAC-Bayesian generalisation error bounds. In addition, Zhang considers various improve-

ments of the theorem, which are required to obtain optimal convergence rates in parametric settings. In Chapter 5 we will discuss a precursor of Theorem 1.4 that was introduced by Barron and Cover [1991]. Grünwald [2007, p. 483] describes its history in minimum description length inference in more detail.

1.3.5 Objective Density Code Lengths

Perhaps the most important insight of MDL theory is that the data compression perspective leads to objective criteria for choosing density code lengths (or, equivalently, a prior π). These code lengths do not represent any prior beliefs, but rather should be interpreted as strategies for data compression. It is worth emphasizing a point made by Grünwald [2007, p. 33]: while a prior belief can be true or false, a strategy cannot be true or false in any sense; it can only be clever or stupid. Let us consider a criterion to measure the cleverness of strategies.

Models Provide a Baseline Let $\mathcal{M} = \{p_1, p_2, \ldots\}$ be a model, and let $L(p)$ be density code lengths relative to this model. Then MDL encodes the data using

$$L_{\text{2-p}}(x^n) = \min_{p \in \mathcal{M}} \left\{ L(p) - \log p(x^n) \right\} \tag{1.14}$$

bits. Consequently, the best code length we could hope to achieve by carefully choosing the density code lengths is to come as close as possible to

$$\inf_{p \in \mathcal{M}} - \log p(x^n) = - \log \hat{p}(x^n), \tag{1.15}$$

where $\hat{p} = \arg\max_{p \in \mathcal{M}} p(x^n)$ denotes the *maximum likelihood* density. Thus (1.15) provides a baseline, against which we may compare our actual code lengths. Although the baseline itself is unachievable (except for degenerate cases, the 'code lengths' (1.15) will not satisfy Kraft's inequality), it turns out that in many cases it is possible to choose $L(p)$ such that the overhead of the MDL code lengths (1.14) compared to the baseline is small.

Finite Models For simplicity, suppose first that $\mathcal{M}$ contains only a finite number m of densities, which is small relative to the sample size

n. (Say, $\log m = O(\log n)$.) Then we can take the uniform prior, leading to constant density code lengths $L(p) = \log m$. As n gets large, $\log m$ bits are negligible compared to $-\log p(x^n)$, which is typically linear in n. This can also be seen in Corollary 1.1, where $\lambda L(p)/n$ goes to zero quickly if $L(p) = O(\log n)$. So the two-part MDL code is almost as good as if we had known the best possible p in advance. Therefore, for sufficiently small models, the uniform prior compresses as well as any other possible prior, and can be chosen for completely objective data compression reasons.

We will see next that surprisingly rich classes $\mathcal{M}$ are still sufficiently small to use a similar objective approach, although the appropriate prior is typically not uniform on $\mathcal{M}$.

Parametric Models Suppose $\mathcal{M} = \{p_\theta \mid \theta \in \Theta \subseteq \mathbb{R}^d\}$ is a *parametric model* that is continuously parametrised by d parameters $\theta = (\theta_1, \ldots, \theta_d)$. Then $|\mathcal{M}| = \infty$ and there exists no code for all possible parameters of $\mathcal{M}$. Nevertheless, we can still apply a similar approach as before if the number of distinguishable distributions in $\mathcal{M}$ (in a sense that will be made more precise later) is sufficiently small.

Suppose that, even though $\mathcal{M}$ itself is infinite, there exist a finite number of densities $p_1, \ldots, p_m$ such that

$$\min_{1 \le j \le m} -\log p_j(x^n) \le -\log \hat{p}(x^n) + C \tag{1.16}$$

for all x^n for some constant C. Then $\mathcal{M}$ can be reduced to a finite model $\ddot{\mathcal{M}} = \{p_1, \ldots, p_m\}$ essentially without harming the compression that can be achieved by its elements, and if m is sufficiently small, we can use the same approach as in the previous section. The density code lengths are then determined, not by the size of $\mathcal{M}$, but by the number of elements in $\ddot{\mathcal{M}}$. This leads to an important insight: the complexity of densities is not an inherent property of the densities themselves, but rather of the smallest number m such that (1.16) can be satisfied, which is a measure for the richness or complexity of $\mathcal{M}$. This approach is similar to the use of *sieves* by Grenander [1981]. Consider the following example.

Example 1.4 (Bernoulli Model)**.** Let $\mathcal{M} = \{p_\theta \mid \theta \in [0,1]\}$ be the Bernoulli model, where $\mathcal{X} = \{0,1\}$ and $p_\theta(1) = \theta$. Distributions are extended to multiple outcomes by taking products: $p_\theta(x^n) = \prod_{i=1}^n p_\theta(x_i)$.

Let $\hat{\theta} = \hat{\theta}(x^n) = \sum_{i=1}^n x_i/n$ denote the maximum likelihood parameter. Lemma 3.4 in Chapter 3 shows that there exists a finite set of parameters $\{\ddot{\theta}_1, \ldots, \ddot{\theta}_m\}$ with $m = O(\sqrt{n})$ such that

$$\min_{1\leq j\leq m} -\log p_{\ddot{\theta}_j}(x^n) \leq -\log p_{\hat{\theta}}(x^n) + C \qquad \text{(for all } x^n\text{)}$$

for some constant C that does not depend on n or x^n. (These points are essentially spaced uniformly in a parametrisation by $\phi = \arcsin\sqrt{\theta}$.) Then the MDL estimator relative to model $\ddot{\mathcal{M}} = \{p_{\ddot{\theta}_1}, \ldots, p_{\ddot{\theta}_m}\}$ with uniform density code lengths $L(j) = \log m$, encodes the data using

$$L_{2\text{-p}}(x^n) = \log m + \min_{1\leq j\leq m} -\log p_{\ddot{\theta}_j}(x^n) \leq \frac{1}{2}\log n + C' - \log p_{\hat{\theta}}(x^n) \tag{1.17}$$

bits, where $\hat{\theta} \in [0,1]$ denotes the maximum likelihood parameter in the full Bernoulli model and C' is a constant dependent on C and the maximum ratio between m and $\sqrt{n}$. In addition, Theorem 1.4 shows that if the data are generated by any Bernoulli distribution with parameter θ, then the λ-MDL estimates relative to the restricted set $\ddot{\mathcal{M}}$ converge to θ at rate

$$\mathbf{E}\, D_\alpha(p_\theta \| \ddot{p}_\lambda) \leq \frac{\lambda \log m + C}{n} \leq \frac{\frac{\lambda}{2}\log n + C'}{n},$$

even if θ is not among the discretised parameters $\{\ddot{\theta}_1, \ldots, \ddot{\theta}_m\}$. (Some readers may notice that this is a $\log n$ factor short of the optimal rate $O(1/n)$; this extra factor can be removed by using a more refined version of Theorem 1.4 [Zhang, 2006].)

For general parametric models it may not always be possible to satisfy (1.16) uniformly for all data $x^n \in \mathcal{X}^n$, but we can come close: let $\Gamma \subset \Theta$ be an arbitrary compact subset of the interior of the parameter space Θ. Then if the maximum likelihood estimator satisfies the central limit theorem for $\theta \in \Gamma$ and $\mathcal{M}$ satisfies certain weak smoothness conditions, (1.16) can be satisfied with the same constant C for all sequences x^n such that the maximum likelihood parameter $\hat{\theta}(x^n)$ lies in Γ [Rissanen, 1996, Grünwald, 2007, Theorem 10.1]. (The latter reference restricts attention to so-called exponential families, but presumably generalises to general parametric families.) In such cases one may approximate Θ by a sequence $\Gamma_1 \subset \Gamma_2 \subset \cdots \subset \Theta$ such that $\bigcup_k \Gamma_k = \Theta$ and for any data x^n let C in (1.16) depend on the smallest k such that $\hat{\theta}(x^n) \in \Gamma_k$. Then

C will be bigger as $\hat{\theta}(x^n)$ lies closer to the boundary of Θ. If the data are sampled from a density $p_\theta \in \mathcal{M}$, then by the law of large numbers there will almost surely be a fixed k such that $\hat{\theta} \in \Gamma_k$ for all sufficiently large n.

Example 1.5 (Normal Location Family)**.** Let $\mathcal{M} = \{p_\mu \mid \mu \in \mathbb{R}\}$ denote the normal location family with densities

$$p_\mu(x) = \tfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

which share a fixed variance σ^2. Densities are again extended to multiple outcomes by taking products. Suppose that the maximum likelihood parameter $\hat{\mu} = \hat{\mu}(x^n) = \sum_{i=1}^n x_i/n$ lies inside a known interval $\Gamma = (a, b)$:

$$a < \hat{\mu}(x^n) < b. \tag{1.18}$$

Then using

$$-\ln p_\mu(x^n) - [-\ln p_{\hat{\mu}}(x^n)] = \frac{n(\mu - \hat{\mu})^2}{2\sigma^2},$$

where ln denotes the natural logarithm, we can cover Γ by $m = (b - a)\sqrt{n}/(2\sigma)$ bins of size $2\sigma/\sqrt{n}$, which ensure that for any x^n satisfying (1.18)

$$\min_{1 \le j \le m} -\ln p_{\ddot{\mu}_j}(x^n) \le -\ln p_{\hat{\mu}}(x^n) + \frac{1}{2},$$

where $\ddot{\mu}_1, \ldots, \ddot{\mu}_m$ denote the centers of the bins. It follows that there exists a two-part code with code lengths (measured in nats)

$$L_\Gamma(x^n) = \ln m + \min_{1 \le j \le m} -\ln p_{\ddot{\mu}_j}(x^n)$$

such that

$$L_\Gamma(x^n) - [-\ln p_{\hat{\mu}}(x^n)] \le \ln m + \frac{1}{2} = \frac{1}{2}\ln n + \ln\frac{b-a}{2\sigma} + \frac{1}{2}$$

for all x^n such that $a < \hat{\mu}(x^n) < b$. Grünwald [2007, Chapter 11] discusses codes with code lengths L such that $L(x^n) - [-\ln p_{\hat{\mu}}(x^n)]$ increases in a principled way as $(a, b) \to (-\infty, \infty)$.

Optimal Discretisation The previous discussion leads to the following question: for general models, what is the smallest number of densities m such that (1.16) can be satisfied? In particular, are $m = O(\sqrt{n})$ for the Bernoulli model and $m = (b-a)\sqrt{n}/(2\sigma)$ for the normal location family with restricted maximum likelihood optimal? These questions may be answered by comparing the two-part code lengths

$$L_{\text{2-p}}(x^n) = \min_{p \in \ddot{\mathcal{M}}} \left\{ L(p) - \log p(x^n) \right\} \tag{1.19}$$

of the MDL code for the discretised model $\ddot{\mathcal{M}}$ to the code lengths of the code that comes as close as possible to the baseline set by $\mathcal{M}$, uniformly on all data $x^n \in \mathcal{Y} \subseteq \mathcal{X}^n$. For example, $\mathcal{Y}$ may be the set of all x^n with maximum likelihood parameter in Γ. Such a code may be a *one-part* code, meaning that it is not required to explicitly encode a density from $\mathcal{M}$, but only needs to encode the data.

For any code with code lengths $L(x^n)$, be it one-part or two-part,

$$R(\mathcal{M}, L, x^n) = L(x^n) - \inf_{p \in \mathcal{M}} -\log p(x^n)$$

is called the *regret* of L with respect to model $\mathcal{M}$ on data x^n, because it measures how much longer our description of the data is when we use L instead of the optimal code based on $\mathcal{M}$ in hindsight, after seeing the data. The *worst-case regret* for $x^n \in \mathcal{Y}$ is therefore

$$\sup_{x^n \in \mathcal{Y}} R(\mathcal{M}, L, x^n).$$

It is uniquely minimized by the code corresponding to the *normalized maximum likelihood* (NML) distribution with density

$$p_{\text{NML}}(x^n) = \frac{\sup_{p \in \mathcal{M}} p(x^n)}{Z(n)},$$

where the normalization

$$Z(n) = \sum_{x^n \in \mathcal{Y}} \sup_{p \in \mathcal{M}} p(x^n)$$

is called the *Shtarkov sum* (relative to $\mathcal{Y}$) [Shtar'kov, 1987] and its logarithm, $\log Z(n)$, is called the *parametric complexity* of $\mathcal{M}$ [Barron et al.,

1998]. For uncountable sample spaces, the sum in the normalization is replaced by an integral over $\mathcal{Y}$. Note that the NML distribution need not exist, because $Z(n)$ may be infinite.

One may view the parametric complexity as the value of a zero-sum game, in which a Statistician first picks a code with the intent of minimizing the regret and then Nature picks the data that maximizes the regret. From this perspective, the code lengths $L_{\text{NML}}(x^n) = -\log p_{\text{NML}}(x^n)$ form an equalizer strategy for Statistician, which achieves the same regret $R(\mathcal{M}, L_{\text{NML}}, x^n) = \log Z(n)$ on all data $x^n \in \mathcal{Y}$. Thus

$$\sup_{x^n \in \mathcal{Y}} R(\mathcal{M}, L_{\text{NML}}, x^n) = \log Z(n).$$

The fact that $P_{\text{NML}}(\mathcal{Y}) = 1$ implies that no other distribution has at least as high density on all $x^n \in \mathcal{Y}$. In terms of codes this means that any other code must have higher regret than L_{NML} for some data $x^n \in \mathcal{Y}$. Therefore the smallest worst-case regret on $\mathcal{Y}$ we could hope for with any code is the parametric complexity, $\log Z(n)$. This holds in particular for the two-part MDL code with code lengths $L_{2\text{-p}}$ as in (1.19).

The Shtarkov sum $Z(n)$ can be interpreted as a volume that is proportional to the number of *distinguishable distributions* in $\mathcal{M}$, where distinguishability is measured using Kullback-Leibler divergence [Grünwald, 2007, Balasubramanian, 1997]. This makes the Shtarkov sum and its logarithm, the parametric complexity, inherent measures of the complexity of $\mathcal{M}$. As a direct consequence, they play a fundamental role in MDL model selection, which is discussed in Section 1.4.

If $\mathcal{Y}$ is the set of all x^n with maximum likelihood parameter in a compact subset Γ of the interior of the parameter space Θ, then the parametric complexity can often be approximated by

$$\log Z(n) = \frac{d}{2} \log \frac{n}{2\pi} + \log \int_{\Gamma} \sqrt{|I(\theta)|}\, \mathrm{d}\theta + o(1), \tag{1.20}$$

where $|I(\theta)|$ denotes the Fisher information at θ, d denotes the number of free parameters of the model and $o(1) \to 0$ as $n \to \infty$. This approximation holds if the maximum likelihood estimator satisfies the central limit theorem for $\theta \in \Gamma$ and $\mathcal{M}$ satisfies certain weak smoothness conditions (which do not require it to be i.i.d.) [Rissanen, 1996]. Although for some models the $o(1)$ term may go to zero very slowly, the approximation is quite accurate in the following examples.

Example 1.4 (cont.) Recall that in the Bernoulli example we have constructed a two-part code with code lengths $L_{2\text{-p}}$ such that the worst-case regret for any data $x^n \in \mathcal{X}^n$ was bounded by $\frac{1}{2}\log n + C'$. In this case the worst-case regret is

$$\log Z(n) = \frac{1}{2}\log n + \frac{1}{2}\log\frac{\pi}{2} + o(1)$$

[Xie and Barron, 2000]. Thus we see that the two-part code with $m = O(\sqrt{n})$ is indeed optimal up to a constant. As $d = 1$ and the Fisher information equals $\theta^{-1}(1-\theta)^{-1}$ for the Bernoulli model, such that

$$\int_0^1 \sqrt{|I(\theta)|}\,\mathrm{d}\theta = \pi,$$

we also see that the approximation (1.20) applies.

Example 1.5 (cont.) In the normal location family the worst-case regret of L_Γ over data in $\mathcal{Y} = \{x^n \mid \hat{\mu}(x^n) \in \Gamma = (a,b)\}$ was bounded by $\frac{1}{2}\log n + \log\frac{b-a}{2\sigma} + \frac{1}{2}$. In this case the parametric complexity is exactly

$$\log Z(n) = \frac{1}{2}\log n + \log\frac{b-a}{\sqrt{2\pi}\sigma}$$

[Grünwald, 2007, p. 298]. As the difference, $\frac{1}{2} + \frac{1}{2}\log\frac{\pi}{2}$, is constant, we find that the two-part code L_Γ is essentially optimal.

As $d = 1$ and the Fisher information is $1/\sigma^2$, it turns out that in this special case the approximation (1.20) is exact with $o(1) = 0$. We also see that $\log Z(n) \to \infty$ as $(a,b) \to (-\infty,\infty)$. Therefore there does not exist a code that achieves finite worst-case regret relative to the unrestricted set $\mathcal{Y} = \mathcal{X}^n$.

1.4 MDL Model Selection

We turn to MDL model selection. Suppose we have a countable number of parametric models $\mathcal{M}_1, \mathcal{M}_2, \ldots$, and the goal is to select one of them based on data x^n, and possibly estimate its parameters as well. For example, $\mathcal{M}_k$ might be the set of Markov chains that take into account the k previous outcomes, parametrised by their transition probabilities. Or $\mathcal{M}_k$ may be the set of all histograms with k fixed-width bins, parametrised by the density in each bin.

As MDL automatically protects against overfitting (Theorems 1.3 and 1.4), we can in principle just define the *model class* $\mathcal{M} = \bigcup_k \mathcal{M}_k$ and use MDL to estimate its parameters. But there is a complication: as this meta-model $\mathcal{M}$ will usually have very large or infinite parametric complexity, we cannot choose a prior that guarantees the same regret on all possible data sequences, as suggested in the previous section. We will therefore have to make more use of the structure supplied by the models and aim at achieving uniform regret only relative to each submodel $\mathcal{M}_k$. This works as follows.

1.4.1 Estimating Both Structure and Parameters

Let $\log Z_k(n)$ be the parametric complexity of $\mathcal{M}_k$ at sample size n, and assume that the models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ are ordered from simple to complex: $Z_1(n) \leq Z_2(n) \leq \cdots$. Let us first consider how to estimate both the model structure k and the parameters of $\mathcal{M}_k$ at the same time. In this case we treat each model $\mathcal{M}_k$ as before, and associate with it a discretised model $\ddot{\mathcal{M}}_k$ and density code lengths L_k for the elements of $\ddot{\mathcal{M}}_k$. We now associate density code lengths

$$L(p) = \min_{\{k:\, p \in \ddot{\mathcal{M}}_k\}} \Big\{ L(k) + L_k(p) \Big\}$$

with the elements of the model class $\ddot{\mathcal{M}} = \bigcup_k \ddot{\mathcal{M}}_k$, where $L(k)$ are code lengths for k that increase only slowly with k. For concreteness, let us take $L(k) = -\log \pi(k)$ with $\pi(k) = k^{-1}(k+1)^{-1}$, such that $L(k) \leq 2\log(k+1)$ increases only logarithmically in k. The fact that $L(k)$ assigns larger code length to models with a higher index k is usually of no concern, since such models also have bigger parametric complexity, so that $L_k(p)$ will typically dominate $L(k)$ anyway. To summarize, model structure k and parameters can be estimated simultaneously using the MDL estimator:

$$\begin{aligned} \ddot{p} &= \arg\min_{p \in \ddot{\mathcal{M}}} \Big\{ L(p) - \log p(x^n) \Big\} \\ &= \arg\min_{k \in \mathbb{N}, p \in \ddot{\mathcal{M}}_k} \Big\{ L(k) + L_k(p) - \log p(x^n) \Big\}, \end{aligned}$$

where we have identified $p \in \ddot{\mathcal{M}}_k$ with the pair (k, p).

Suppose the models are i.i.d. Then by Corollary 1.1 the λ-MDL estimator $\ddot{p}_\lambda$ converges at rate

$$\mathbf{E}\, D_\alpha(q\|\ddot{p}_\lambda) \leq \min_{k\in\mathbb{N}, p\in\ddot{\mathcal{M}}_k} \left\{ \frac{2\lambda \log(k+1) + \lambda L_k(p)}{n} + D(q\|p) \right\}$$

to any i.i.d. density q. Note that q need not lie in any of the models, as long as it can be approximated in Kullback-Leibler divergence by a sequence of elements from $\mathcal{M}$. This is the case, for example, in density estimation with histograms, where the estimated densities are typically not histograms themselves.

1.4.2 Estimating Structure Only

Suppose we are interested only in selecting a single model among the candidates $\mathcal{M}_1, \mathcal{M}_2, \ldots$ and we do not need to estimate the parameters of the model at the same time. This may be the case, for example, in linear regression if each of the models corresponds to a different subset of the regressor variables. The reason for model selection may then be to determine the relevant variables, while an estimate of their coefficients is not required [Grünwald, 2007, p. 25]. In such a setting, instead of explicitly discretising each $\mathcal{M}_k$ into $\ddot{\mathcal{M}}_k$ such that $L_k(x^n) = \min_{p\in\ddot{\mathcal{M}}_k} L_k(p) - \log p(x^n)$ achieves worst-case regret close to the parametric complexity, we can directly use the normalised maximum likelihood code! This gives rise to the following MDL model selection procedure:

$$\arg\min_k L(k) - \log p_{\mathrm{NML},k}(x^n) = \arg\min_k L(k) + \log Z_k(n) - \log \hat{p}_k(x^n),$$

where $p_{\mathrm{NML},k}$ and $\hat{p}_k$ denote the normalised maximum likelihood (NML) density and the maximum likelihood density for model $\mathcal{M}_k$, respectively. This procedure is *optimal* in the sense that $p_{\mathrm{NML},k}$ achieves the smallest possible worst-case regret relative to $\mathcal{M}_k$.

We see that MDL model selection may be interpreted as a penalised maximum likelihood procedure, which penalises model $\mathcal{M}_k$ by its parametric complexity $\log Z_k(n)$ (and $L(k)$, but the influence of this term is usually small). The parametric complexity arises unavoidably from coding considerations as the smallest possible worst-case regret relative to model $\mathcal{M}_k$, and $Z_k(n)$ has an interpretation as the number of distinguishable distributions in $\mathcal{M}_k$. Thus, unlike the ordinary maximum

likelihood estimator, the MDL estimator scales up from parameter estimation to model selection without modification, and its complexity penalty does not arise from asymptotic analysis under probabilistic assumptions, but has a coding interpretation at the actual sample size n.

Remark 1.3 (Luckiness). One may also look at model selection the other way around. Suppose we have a big model $\mathcal{M}$ with (too) large parametric complexity compared to the sample size. For example, we may have $\log Z(n) \approx n$. Then there is no point in treating all elements of $\mathcal{M}$ on the same footing, since their estimate will not converge (see Theorem 1.4). In such cases it is essential to introduce more bias into the choice of the density code lengths. This can be done by carving up $\mathcal{M}$ into submodels $\mathcal{M}_1, \mathcal{M}_2, \ldots$ such that $\mathcal{M} = \bigcup_k \mathcal{M}_k$. One may then proceed as above, regarding $\mathcal{M}_1, \mathcal{M}_2, \ldots$ as models and $\mathcal{M}$ as a model class that is their union. This is called a *luckiness* approach: if one is lucky, the data can be compressed by a submodel with small parametric complexity and their structure will be learned at a small sample size; if one is unlucky then one will have added $L(k) + \log Z_k(n) - \log Z(n) \leq L(k)$ bits to the density code lengths, which is typically small compared to $\log Z(n)$, so one will not lose much. The use of luckiness is advocated by De Rooij and Grünwald [2010].

1.4.3 Universal Coding

In practice MDL model selection is often based on approximations of the normalised maximum likelihood density. There are two options here: one is to use (1.20), but that only really works if the $o(1)$ term it contains is sufficiently small. Otherwise none of MDL theory applies and we do not have any guarantees about performance. A better alternative is therefore to apply MDL with distributions that approximate the NML distribution. Although such distributions may not exactly minimize the worst-case regret, at least they define real codes and therefore the data compression interpretation and theoretical results still apply. One example is the two-part code above, based on discretising each $\mathcal{M}_k$ into a corresponding $\ddot{\mathcal{M}}_k$. But there are two important other choices as well. Such codes, which try to minimize the worst-case regret compared to $\mathcal{M}_k$, will be called *universal codes*. This informal definition is slightly stronger than the standard definition [Grünwald,

2007]. Under our definition, the NML code is the optimal universal code.

1.4.3.1 Bayesian Universal Code

There is a close connection between MDL and Bayesian model selection. The reason is that, for appropriate priors, the Bayesian marginal likelihood often achieves small worst-case regret, and sometimes even gets close to the optimal worst-case regret: the parametric complexity.

Countable Models Suppose $\mathcal{M} = \{p_1, p_2, \ldots\}$ is a countable model and presume, for simplicity, that the data are discrete. Then, given a prior probability mass function w on $\mathcal{M}$, the *Bayesian marginal likelihood* is the distribution defined by

$$b(x^n) = \sum_{p \in \mathcal{M}} w(p) p(x^n). \tag{1.21}$$

The corresponding code with code lengths $-\log b(x^n)$ is called the *Bayesian universal code*. Let us compare this to the two-part code

$$L_{\text{2-p}}(x^n) = \min_{p \in \mathcal{M}} L(p) - \log p(x^n)$$

with density code lengths $L(p) = -\log w(p)$, which corresponds to the (incomplete) distribution defined by

$$p_{\text{2-p}}(x^n) = \max_{p \in \mathcal{M}} w(p) p(x^n).$$

As the sum in (1.21) can be bounded from below by its largest term, we find that $b(x^n) \geq p_{\text{2-p}}(x^n)$ and therefore the Bayesian universal code always achieves shorter code lengths than the two-part universal code:

$$-\log b(x^n) \leq L_{\text{2-p}}(x^n) \qquad \text{for all } x^n.$$

It follows that, from a data compression point of view, the Bayesian marginal likelihood should always be preferred over the two-part code based on the same prior! Unfortunately, the Bayesian universal code cannot be applied directly to parameter estimation, but it can be used in model selection. MDL model selection then becomes

$$\min_k L(k) - \log b_k(x^n),$$

where b_k denotes the Bayesian marginal likelihood for $\mathcal{M}_k$. For the prior $\pi(k)$ such that $L(k) = -\log \pi(k)$, we see that a model $\mathcal{M}_k$ is preferred over another model $\mathcal{M}_m$ if

$$\frac{b_k(x^n)}{b_m(x^n)} > \frac{\pi(m)}{\pi(k)}.$$

The left-hand side of this expression may be recognised as the *Bayes factor* [Kass and Raftery, 1995]. Bayes factors model selection is therefore formally equivalent to MDL model selection using Bayesian universal codes. As will be seen next, this correspondence generalises to continuous sample spaces and uncountable models, as long as within-model priors w_k are used that ensure the Bayesian universal code has small worst-case regret.

Parametric Models Let $\mathcal{X} = \mathbb{R}^a$ for any finite dimension a, and let

$$\mathcal{M} = \{p_\theta \mid \theta \in \Theta \subseteq \mathbb{R}^d\} \tag{1.22}$$

be a parametric model with d parameters, with densities extended to multiple outcomes by taking products. Then for any prior density w on Θ, the density of the Bayesian marginal likelihood becomes

$$b(x^n) = \int w(\theta) p_\theta(x^n) \, \mathrm{d}\theta. \tag{1.23}$$

With a suitable choice of w, this turns out to be a good candidate to approximate the NML distribution. For a class of models called *exponential families* [Grünwald, 2007] it is even possible to (asymptotically) achieve the optimal worst-case regret, equal to (1.20). The class of exponential families includes the Poisson family, the geometric family of distributions, and the Bernoulli and multinomial models. Furthermore, the set of normal distributions of arbitrary mean and variance is also an exponential family. On the other hand, there are also many parametric models that are not exponential families, like mixtures of normal distributions.

Suppose that $\mathcal{M}$ is an exponential family. Then for data x^n such that the maximum likelihood parameter $\hat{\theta}$ lies is a compact subset of the interior of Θ, and a continuous prior density w that is bounded away

from zero, the regret of the Bayesian universal code can asymptotically be approximated by

$$-\log b(x^n) - [-\log p_{\hat{\theta}}(x^n)] = \frac{d}{2}\log\frac{n}{2\pi} + \log\frac{\sqrt{|I(\hat{\theta})|}}{w(\hat{\theta})} + o(1), \quad (1.24)$$

where $|I(\theta)|$ denotes the Fisher information at θ [Grünwald, 2007, Theorem 8.1]. Comparison with (1.20) shows that Bayes asymptotically achieves the parametric complexity if we use *Jeffreys' prior*

$$w(\theta) = \frac{\sqrt{|I(\theta)}}{\int \sqrt{|I(\theta)}\,\mathrm{d}\theta}.$$

Other priors are also acceptable as long as they dominate Jeffreys' prior. Grünwald [2007, Chapter 8] provides a discussion of the extent to which these results extend beyond exponential families.

The preceding discussion provides a data compression motivation for the use of Bayesian universal codes. There also exists a frequentist motivation, which holds for arbitrary parametric models, not just exponential families. This frequentist motivation is that MDL model selection with the Bayesian marginal likelihood is *consistent*, in the sense that *if* the data are sampled from a distribution in one of the models, then it selects that model with probability one for all sufficiently large samples [Dawid, 1992b, Barron et al., 1998].

The formal statement of this result (Theorem 1.5 below) requires an interpretation of the Bayesian marginal likelihood as a probabilistic source B^∞ on infinite sequences of outcomes $x^\infty = x_1, x_2, \ldots$ This source has the marginal distribution defined by (1.23) for any finite number of outcomes n. For two parametric models $\mathcal{M}_1$ and $\mathcal{M}_2$, the corresponding sources B_1^∞ and B_2^∞ will often be quite different, in the sense that there exists a measurable event $A \subseteq \mathcal{X}^\infty$ such that $B_1^\infty(A) = 1$ and $B_2^\infty(\mathcal{X}^\infty \setminus A) = 1$. In this case B_1^∞ and B_2^∞ are called *mutually singular*. Mutual singularity is quite common. It occurs, for example, if the models contain stationary ergodic distributions and the priors are mutually singular on the space of distributions. This is the case, for example, if the models are parametric families of i.i.d. or Markov distributions, and the parameter spaces are of different dimensionality and absolutely continuous prior densities are assigned to each dimension [Barron et al., 1998, Dawid, 1992b]. See Section 2.6 of Chapter 2 for further discussion.

Section 6.5.7 of Chapter 6 also relates mutual singularity on infinite sequences to Rényi divergence.

Theorem 1.5 (Model Selection Consistency)**.** *Let $\mathcal{M}_1, \mathcal{M}_2, \ldots$ be parametric models of form (1.22) with priors $w_1, w_2, \ldots$ such that the corresponding Bayesian marginal likelihoods $B_1^\infty, B_2^\infty, \ldots$ on infinite sequences are mutually singular. Let Θ_k denote the parameter space of $\mathcal{M}_k$ and let*

$$\ddot{k} = \ddot{k}(X^n) = \arg\min_k L(k) - \log b_k(X^n)$$

denote the MDL estimator for model selection with Bayesian universal codes on a sample $X_1, \ldots, X_n$ of size n, where the code lengths $L(k)$ are finite for all k. Then, for all k^*, for w_{k^*}-almost all $\theta^* \in \Theta_{k^*}$,

$$\ddot{k} = k^*$$

for all sufficiently large n, with P_{θ^*}-probability one.

When the asymptotic expansions (1.20) and (1.24) for NML and Bayes with Jeffreys' prior hold uniformly for all models, Theorem 1.5 implies that model selection based on NML is also consistent. However, whether this is typically the case is not known.

Predictive Interpretation When models are used to make predictions, it is often convenient to look at the Bayesian and NML universal codes in a different way: their code lengths can be interpreted as the cumulative loss incurred when sequentially predicting the data $x^n = x_1, \ldots, x_n$, and as a consequence, MDL and Bayes factors model selection can be interpreted as selecting the model with smallest cumulative prediction error.

For simplicity, assume that $\mathcal{X}$ is countable. Consider the following sequential prediction problem: for $t = 1, \ldots, n$, predict x_t given knowledge only of the preceding outcomes $x_1, \ldots, x_{t-1}$ by specifying a probability distribution P_t for x_t. The quality of predictions is measured by the *log(arithmic) loss*:

$$\ell(x_t, P_t) = -\log P_t(x_t),$$

which may be interpreted as the code length of x_t under the code corresponding to P_t. Suppose that Q is any distribution on all data x^n and

P_t is its marginal distribution on x_t conditioned on the preceding outcomes x^{t-1}, so that $P_t(x_t) = Q(x_t \mid x^{t-1})$. Then the cumulative loss on x^n is

$$\sum_{t=1}^{n} \ell(x_t, P_t) = -\log \prod_{t=1}^{n} P_t(x_t) = -\log Q(x^n).$$

We see that the cumulative prediction error of predicting according to the conditional distributions of Q is equal to the code length of x^n under Q. In particular, by letting Q be the Bayesian marginal likelihood, we can rewrite MDL or Bayes factors as selecting the model that achieves

$$\min_k L(k) + \sum_{t=1}^{n} \ell(x_t, Q(\cdot \mid x^{t-1})).$$

Thus, apart from a constant offset $L(k)$ per model, MDL or Bayes factors selects the model with smallest cumulative prediction error when sequentially predicting the data. This interpretation is especially appropriate if the selected model is to be used to make predictions of future data (see Chapter 2): in the case of the Bayesian marginal likelihood, the prediction of a hypothetical new outcome x_{n+1} outside of the given sample would be $Q(x_{n+1} \mid x^n)$. We see that the corresponding loss $\ell(x_{n+1}, Q(\cdot \mid x^n))$ is just a continuation of the sequence of losses on the sample x^n.

1.4.3.2 Plug-in Universal Code

In the previous section we saw that the conditional distributions of a universal code can be used as sequential predictions P_t. Reversing the construction, one can also plug in the predictions of some estimator for the parameters of a parametric model $\mathcal{M} = \{p_\theta \mid \theta \in \Theta \subseteq \mathbb{R}^d\}$ as conditional probabilities to construct a universal code. For any estimator $\hat{\theta} \colon \mathcal{X}^{t-1} \to \Theta$ that is defined for all t, this gives:

$$P_{\text{pl}}(x^n) = \prod_{t=1}^{n} P_{\hat{\theta}(x^{t-1})}(x_t \mid x^{t-1}).$$

The fact that the estimator used to predict x_t depends only on the preceding outcomes x^{t-1} and not on x_t or any following data, guarantees that $-\log P_{\text{pl}}(x^n)$ is a code length. The corresponding code, with code lengths $-\log P_{\text{pl}}(x^n)$, is called a *plug-in* code. Thus, letting $P^1_{\text{pl}}, P^2_{\text{pl}}, \ldots$ be

plug-in codes for models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ with estimators $\hat{\theta}_1, \hat{\theta}_2, \ldots$, MDL selects the model that achieves

$$\min_k L(k) - \log P_{\text{pl}}^k(x^n) = \min_k L(k) + \sum_{t=1}^{n} -\log P_{\hat{\theta}_k(x^{t-1})}(x_t \mid x^{t-1}).$$

If a (possibly smoothed, see below) maximum likelihood estimator is plugged in, then MDL model selection based on the plug-in universal code is often *consistent*: if the data are sampled from a distribution in one of the models, then it selects that model with probability one for all sufficiently large samples [Dawid, 1992b, de Luna and Skouras, 2003, Hemerly and Davis, 1989]. This is in contrast to the maximum likelihood estimate on all data, $-\log P_{\hat{\theta}(x^n)}(x^n)$, which does not satisfy Kraft's inequality and is susceptible to overfitting when used as a basis for model selection. Plug-in codes appeal when the same estimator $\hat{\theta}$ is used for prediction of future data, outside of the sample x^n. The plug-in code lengths are also sometimes easier to compute than the code lengths for other universal codes, especially if the models are not i.i.d.

Theoretical analysis of plug-in codes has focused on smoothed versions of the maximum likelihood estimator. The following example illustrates smoothing and its necessity.

Example 1.4 (cont.) Let $\mathcal{M} = \{p_\theta \mid \theta \in [0,1]\}$ be the Bernoulli model, and let $\hat{\theta} = \hat{\theta}(x^n) = n_1/n$ denote the ordinary maximum likelihood estimator, where n_y denotes the number of occurrences of y in x^n. Although the maximum likelihood estimator is consistent if the data are generated by a Bernoulli distribution, its estimates are often extreme on very small samples, which ruins its coding performance. To illustrate, consider data x^n such that the first three outcomes are $0, 0, 1$. Then $P_{\hat{\theta}(x^2)}(x_3) = 0$, such that

$$-\log P_{\text{pl}}(x^n) \geq -\log P_{\hat{\theta}(x^2)}(x_3) = \infty.$$

This problem can be avoided by changing the estimator to

$$\hat{\theta}'(x^n) = \frac{n_1 + a}{n + a + b}$$

for positive numbers a and b. These numbers may be interpreted as adding a fake ones to the data and b fake zeroes, such that every possible outcome has been observed before seeing the real data. This prevents the estimator from assigning probability zero to any outcome.

In general, padding the data with initial fake outcomes to prevent the maximum likelihood estimator from giving zero probability is called *smoothing*.

Under certain conditions, including that the data are sampled from an element of the model, the plug-in code P_{pl} based on a smoothed maximum likelihood estimator is guaranteed to have small regret in a weak expected sense:

$$\mathbf{E}\left[-\log P_{\text{pl}}(X^n) + \log P_{\hat{\theta}(X^n)}(X^n)\right] = \frac{d}{2}\log n + O(1), \tag{1.25}$$

where the expectation is with respect to an element of $\mathcal{M}$ [Grünwald, 2007] and the term $O(1)$ can asymptotically be bounded above by a constant. See also [Rissanen, 1986, 1989, Wei, 1992]. As in model selection it is usually not the case that all models contain the true distribution, (1.25) does not directly justify using plug-in codes based on smoothed maximum likelihood estimators for model selection. Indeed, although model selection based on the plug-in code for a smoothed maximum likelihood estimator is typically consistent (see above), it performs somewhat worse than model selection based on NML or Bayes [Grünwald and de Rooij, 2005]. However, it has recently been shown that it is possible to construct different estimators that do not suffer from this problem [Kotłowski et al., 2010, Grünwald and Kotłowski, 2010].

1.4.4 Nonparametric Models

MDL may also be used in nonparametric settings. Barron and Cover [1991] work out two applications, in which they show that the right-hand side of (1.12) in Corollary 1.1, which they call the *index of resolvability*, converges to zero at a certain rate. The first application follows the pattern of Section 1.4.1: the models are either polynomials or splines of order k. They construct a two-part universal code for every model and also encode k to allow the appropriate order to be determined automatically. Then they show that the index of resolvability converges to zero at rate $O((\frac{\log n}{n})^{2r/(2r+1)})$ if the data are drawn i.i.d. from any density p on the open interval $(0,1)$ that satisfies the smoothness condition

$$\int_0^1 \left(\frac{\mathrm{d}^r}{\mathrm{d}x^r}\log p(x)\right)^2 \mathrm{d}x < \infty$$

for some unknown $r \geq 1$. Notably, the indicated rate of convergence holds without prior knowledge of r. Note also that the densities satisfying the smoothness condition need not be polynomials or splines themselves, as long as they can be arbitrarily well approximated in Kullback-Leibler divergence by polynomials or splines, respectively.

Barron and Cover also provide a fully nonparametric example, which does not involve any parametric models at all. In this example a nonparametric set of densities is reduced to a finite set in a way similar to the construction for parametric models in Section 1.3.5. The size of the finite set is determined by the so-called Kolmogorov ε-entropy of the nonparametric set of densities. See [Barron and Cover, 1991] for details.

While the results of Barron and Cover apply to two-part codes, other work has focused on other universal codes. Notably, Seeger et al. [2008] prove a regret bound for Bayesian *Gaussian process* models in nonparametric regression. They show that the regret compared to a regression function from a reproducing kernel Hilbert space that is determined by the parameters of the Gaussian process, grows quadratically with the norm of the regression function. It follows that Gaussian processes have very good universal coding properties relative to regression functions with small norm. See also [Grünwald, 2007].

Instead of the regret, Rissanen et al. [1992] and Yu and Speed [1992] analyse the closely related *redundancy*, which may be regarded as an in-expectation analog of the regret. Using histogram models to estimate bounded densities on the unit interval with bounded derivatives, they find that MDL model selection based on Bayesian universal codes misses the optimal minimax redundancy by (only) a logarithmic factor. In Chapter 2 we provide an explanation for this lack of optimality, which we call the *catch-up phenomenon*. Based on this explanation, we introduce a new model selection method, which still has a data compression interpretation, but provably does not suffer from the catch-up phenomenon. This method can be applied to arbitrary models, for which the impact of the catch-up phenomenon may be larger.

1.5 Organisation of this Thesis

As we have seen, MDL inference views densities and models as strategies for data compression. This stands in sharp contrast to making

assumptions about an underlying distribution generating the data, as is standard even in nonparametric statistics. Strategies are either good or bad, and certainly we do not expect bad models to magically lead to good inference. But, unlike assumptions, strategies can never be true or false. Therefore, if the MDL premise of making data compression a fundamental notion can hold its ground, it promises a robust kind of statistics, which does not break down when standard, but hard to verify, assumptions fail.

This makes it worthwhile to stress test the data compression principles behind the minimum description length principle. A natural starting point are cases where MDL disagrees with a more standard frequentist analysis. This thesis analyses two such cases. The first case, studied in Part I, deals with switching between prediction strategies. The second case, described in Part II, deals with the strange $\lambda > 1$ condition from Theorem 1.4.

1.5.1 Part I: Switching between Models

In *Chapter 2* it is found that standard MDL model selection, as described in this introduction, may lead to suboptimal predictions of future data. This problem is then remedied by constructing a code that achieves better data compression than the standard code, by combining models into a meta-model that sequentially switches between them. Thus we see that standard MDL fails, but the underlying principle, data compression, holds its ground. *Chapters 3 and 4* study the general phenomenon of switching between predictors from a related perspective called *prediction with expert advice*. In Chapter 3 a new method is introduced that automatically determines the optimal switching rate when switching between predictors. In Chapter 4 we discuss whether the parts between switches should be modelled independently, or as part of the rest of the data. A new method is introduced to deal with the first case, which is appropriate, for example, for certain time series data.

As there are many connections between MDL and other statistical methods, studying MDL usually sheds new light on other methods as well. To bring out these connections, Chapter 2 uses statistical terminology and restricts attention to the Bayesian and plug-in universal codes, as these correspond to widely used Bayesian and frequentist methods. The results of Chapters 3 and 4 on the other hand are described using the framework of prediction with expert advice.

1.5.2 Part II: MDL Convergence and Rényi Divergence

In this introduction multiple motivations for using the MDL estimator have been presented. There is, however, some tension between them. On the one hand the frequentist convergence result Theorem 1.4 suggests that λ-MDL should be used with $\lambda > 1$. But, on the other hand, from the data compression point of view this just seems to be wasting bits and should therefore be avoided. It seems that at least one of the two ideas must be missing something.

Chapter 5 takes a closer look at Theorem 1.4. It explains how $\lambda > 1$ may be interpreted as a condition on the density code lengths, and examples are given that show that ordinary 1-MDL need not converge at all if this condition is completely removed. Although no definitive verdict is reached on the appropriateness of data compression as a fundamental principle, two new theorems comparable to Theorem 1.4 (but using a weaker mode of convergence) are proved, which show that $\lambda > 1$ is actually a stronger condition than necessary. This shows that the $\lambda > 1$ condition does not tell the full story either, and should not be interpreted as a necessary requirement. Some preliminary consequences of the theorems are presented, which do not follow from the $\lambda > 1$ condition.

The new theorems formulate conditions on the density code lengths in terms of Rényi divergence, but although Rényi divergence was introduced in the nineteen-sixties and appears in many computations, there exists no overview of its technical properties. *Chapter 6* remedies this situation by formally proving the basic properties of Rényi divergence.

Part I

Switching between Models

Chapter 2

Catching Up Faster by Switching Sooner:
A predictive approach to adaptive estimation with an application to the AIC-BIC Dilemma

Prediction and estimation based on Bayesian model selection and model averaging, and derived methods such as BIC, do not always converge at the fastest possible rate. We identify the *catch-up phenomenon* as a novel explanation for the slow convergence of Bayesian methods, and use it to define a modification of the Bayesian predictive distribution, called the *switch distribution*. When used as an adaptive estimator, the switch distribution does achieve optimal cumulative risk convergence rates in nonparametric density estimation and Gaussian regression problems. We show that the minimax cumulative risk is obtained under very weak conditions and without knowledge of the underlying degree of smoothness.

Unlike other adaptive model selection procedures such as AIC and leave-one-out cross-validation, BIC and Bayes factor model selection are typically statistically consistent. We show that this property is retained by the switch distribution, which thus solves the AIC-BIC dilemma for cumulative risk. We give a 'prequential' interpretation to the switch distribution, show how to efficiently implement it, and illustrate its performance on a regression problem with simulated data.

2.1 Introduction

Given a countable number of models (sets of probability distributions), we consider the related tasks of *model selection*, *model averaging* and *adaptive estimation*. In model selection, the goal is to find the model that best

explains the given data. In model averaging, one aims to predict future data from the same source based on a weighted combination of the models. The inferred model or model average may further be used as a basis for adaptive density and regression estimation, in which the goal is to construct estimators that are simultaneously minimax rate optimal with respect to different classes of smoothness.

Some broadly applicable model selection methods such as AIC [Akaike, 1974] and leave-one-out cross-validation (LOO) [Stone, 1977] lead to predictions and corresponding adaptive estimators that are risk optimal in a variety of settings. On the other hand, other popular methods such as the BIC criterion [Schwarz, 1978] and related methods such as Bayes factor model selection [Kass and Raftery, 1995], standard minimum description length (MDL) model selection [Barron et al., 1998] and prequential model validation [Dawid, 1984] are typically suboptimal for prediction and estimation: in many settings, at sample size n the convergence of Bayes factors, MDL, and BIC is a factor $O(\log n)$ slower [Rissanen et al., 1992, Foster and George, 1994, Yang, 1999, Grünwald, 2007]. In this chapter we argue that the slow convergence of Bayes factors (and other BIC-like methods) is caused by the *catch-up phenomenon*, which we will introduce shortly. Our attempt to address this problem takes the form of the *switch distribution*, a practical method (an efficient and very simple algorithm is given in Section 2.2.5) that can be used either directly to predict new outcomes sequentially, or as a basis for model selection and adaptive estimation. The switch distribution may be viewed as an extension of Bayesian Model Averaging or Bayes factor model selection. The standard Bayes factor method is based on a prior distribution on a countable set of distributions $p_1, p_2, p_3, \ldots$; usually, but not necessarily, these are themselves Bayesian marginal distributions relative to some parametric models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ In contrast to a prior on $p_1, p_2, \ldots$, the switch distribution employs a prior defined on *sequences* of the $p_1, p_2, \ldots$, allowing different p_j, and thus different models $\mathcal{M}_j$, to be used for prediction at different sample sizes. In our treatment, as explained in Section 2.3, the p_j are viewed as prediction strategies which may be Bayesian marginal distributions but can also be based on estimators such as maximum likelihood or least-squares. In this sense the switch distribution is more general than a Bayesian marginal distribution and is best interpreted as a *prequential forecasting system* [Dawid, 1984].

The general idea behind the switch distribution is explained further in Section 2.1.2. Our first main result, Theorem 2.1 in Section 2.5.3, shows that in a general i.i.d. setting that includes many nonparametric density and Gaussian regression estimation problems, adaptive estimation based on the switch distribution is optimal relative to the *cumulative Kullback-Leibler (KL) risk*. More precisely, suppose that data are sampled from a density p^*, and p^* is estimated based on a collection of parametric models, where the number of considered models is not more than polynomial in the sample size. Then, as long as the problem is not "too easy", unlike for Bayesian model averaging, the ratio of the cumulative risk incurred by the switch distribution and that incurred by any model selection criterion whatsoever converges to 1. By the problem being "not too easy" we mean that the minimax cumulative risk should be at least of order $(\log n)^2$, a requirement that is satisfied for all nonparametric classes including the standard Sobolev, Hölder and Besov classes [Yang and Barron, 1999]. Thus, the switch distribution may be interpreted as an adaptive estimator which achieves minimax rates without knowledge of the underlying degree of smoothness. The proof requires that the switch distribution is defined with respect to an augmented set of prediction strategies, which increases the time required to process a sample of size n by a factor n. As an alternative we provide Theorem 2.2, which is based on a version of the switch distribution that uses only two prediction strategies per considered model, and therefore has a much faster implementation. The drawbacks are that we impose stronger conditions on the considered models, and that the ratio of cumulative risks may converge to a constant larger than 1. In Section 2.7 we provide experiments with simulated data which suggest that both switch distributions also perform well in practice with small samples.

In the statistical literature, predictive performance is usually measured in terms of instantaneous risk rather than cumulative risk. As shown in Proposition 2.3 (Section 2.8.3), under the conditions of the fast switch distribution, both versions of the switch distribution may be further modified so that they achieve the minimax instantaneous KL risk to within a constant factor larger than one.

2.1.1 Main Application: the AIC-BIC Dilemma

Compared to other broadly applicable model selection criteria such as AIC and LOO, the main advantage of the switch distribution is its provable rate optimality under substantially weaker conditions. A second advantage is that, unlike AIC and LOO, the switch distribution is statistically consistent under fairly weak conditions, i.e. the probability under the true distribution that the correct model is selected converges to 1. This is shown in our third main result, Theorem 2.3. Thus, switching resolves a version of the AIC-BIC dilemma where predictive performance is measured in terms of cumulative risk [Yang, 2005, 2007a,b]. This dilemma concerns the question whether in any given practical situation, one should adopt an AIC-type method (close to optimal for prediction, yet inconsistent) or a BIC-type method (suboptimal for prediction, yet consistent): we show that, when one is interested in cumulative risk, then in contrast to AIC, the switch distribution is consistent, and in contrast to BIC, it is rate optimal. In adaptive estimation however, it may often be more appropriate to consider the instantaneous rather than the cumulative risk. In this scenario, a result of Yang [2005] applies, which (roughly) states that in the parametric context, there can be no method that achieves both consistency and a minimax optimal convergence rate. Relating our results to this second interpretation is more subtle; some connections are indicated in the discussion (Section 2.8).

2.1.2 Main Idea: the Catch-Up Phenomenon

Suppose we use parametric models $\mathcal{M}_k = \{p_\theta \mid \theta \in \Theta_k\}$ to describe a sequence of observations $x^n = x_1, \ldots, x_n$, where each outcome is drawn from some space $\mathcal{X}$; for simplicity we assume $\mathcal{X}$ to be countable in this introduction, but we do not have this restriction in the rest of the chapter.

In Bayes factors model selection or Bayesian model averaging, prior densities w_k are defined for the parameter spaces Θ_k of each model $\mathcal{M}_k$. We can subsequently compute the Bayesian marginal likelihood of the data as follows:

$$p_k(x^n) = \int_{\theta \in \Theta_k} p_{k,\theta}(x^n) w_k(\theta) \, \mathrm{d}\theta. \tag{2.1}$$

Additionally, a prior mass function π on the model indices $\{1, 2, \ldots\}$ is defined. The Bayes factors approach to model selection is to select the

model k with maximum posterior probability

$$\pi(k \mid x^n) = \frac{p_k(x^n)\pi(k)}{\sum_{k'} p_{k'}(x^n)\pi(k')}.$$

In prediction, Bayesian model averaging (BMA) proceeds based on the marginal distribution on data $p_{\text{bma}}(x^n) = \sum_k p_k(x^n)\pi(k)$. BMA predicts any new outcome $x_{n+1} \in \mathcal{X}$ outside of the sample x^n according to $p_{\text{bma}}(x_{n+1} \mid x^n)$, which is equal to a combination of the models' predictions in which the models are weighted according to their posterior probability:

$$p_{\text{bma}}(x_{n+1} \mid x^n) = \sum_k p_k(x_{n+1} \mid x^n)\pi(k \mid x^n). \tag{2.2}$$

We now discuss how the predictions $p_{\text{bma}}(x_{n+1} \mid x^n)$ and $p(x_{n+1} \mid x^n)$ may be interpreted as a continuation of predictions on the sample x^n, and how $-\log p_{\text{bma}}(x^n)$ and $-\log p_k(x^n)$ may be interpreted as the cumulative prediction error of p_{bma} and p_k on x^n.

Let p be any distribution on samples x^n, like for example p_k or p_{bma}. Then for most x^n the probability $p(x^n)$ is exponentially small in n. It is therefore common to consider $-\log p(x^n)$, which we call the *code length* of x^n. Note that small code length corresponds to large probability. Here and in the remainder we let log denote the logarithm to base two, so that code length is measured in bits. Our terminology is motivated by the Kraft inequality in information theory, which links code lengths to probability distributions [Cover and Thomas, 1991], but code length may also be interpreted as the cumulative *log(arithmic) loss* incurred when sequentially predicting $x_1, \ldots, x_n$ by conditioning p on the past [Barron et al., 1998, Grünwald, 2007, Dawid, 1984, Rissanen, 1984]. To see this, assume the outcomes $x^n = x_1, \ldots, x_n$ are given in a natural order (if not, pick some order at random), and let $x^i = x_1, \ldots, x_i$ denote the first i of them. The $(i+1)$-th outcome is then predicted by the conditional probability $p(x_{i+1} \mid x^i) = p(x^{i+1})/p(x^i)$, and the quality of this prediction is measured by the log loss $-\log p(x_{i+1} \mid x^i)$. Summing up the prediction errors, we see that the code length of the sample is equal to the cumulative log loss of the predictions:

$$\sum_{i=1}^{n} -\log p(x_i \mid x^{i-1}) = -\log \prod_{i=1}^{n} p(x_i \mid x^{i-1}) = -\log p(x^n). \tag{2.3}$$

In particular, the code lengths of p_{bma} and p_k may be interpreted as cumulative prediction errors on the sample. Furthermore, if we predict an $(n+1)$-st outcome outside of the sample x^n according to $p(x_{n+1} \mid x^n)$, the loss we incur may be viewed as the continuation of the sequence of losses within the sample. (Again, this holds for both p_{bma} and p_k.) As such, the fact that the sample contains n outcomes is not particularly special, and may equivalently be viewed as truncating an infinite sample after the first n observations. From this perspective, it is natural to study what happens when n is varied, even if one is only interested in prediction for any particular n.

Like the prediction $p_{\text{bma}}(x_{n+1} \mid x^n)$, the posterior probability $\pi(k \mid x^n) \propto p_k(x^n)\pi(k)$ may also be interpreted in terms of code length: apart from the constant (i.e. not dependent on n) influence of the prior $\pi(k)$, it assigns large probability to models $\mathcal{M}_k$ that give large probability $p_k(x^n)$ to the data or, equivalently, achieve small code length or cumulative prediction error as measured by log loss. Note that the ratio of posterior probabilities of two models is *exponential* in their difference in code length!

We are now ready to compare the predictive performance of BMA to the best possible predictions based on the models. To this end, let $\hat{k} \equiv \hat{k}(x^n) = \arg\min_k -\log p_k(x^n)$ denote the index of the model achieving the smallest cumulative loss (or code length) when sequentially predicting x^n. Then prediction using BMA guarantees that the difference between our code length and the code length achieved by $\hat{k}$ is in the range $[0, -\log \pi(\hat{k})]$, whatever data x^n are observed. (This follows by (2.3) and bounding the sum $\sum_k p_k(x^n)\pi(k)$ from below by the term for $\hat{k}$ and from above by $p_{\hat{k}}(x^n)$.) If, for all k, $-\log \pi(k)$ (which is constant in n) is small compared to $-\log p_k(x^n)$ (which is typically linear in n), then this implies that BMA predicts essentially as well as the model that turns out to be the best one in retrospect, whatever this model may be. Although this is quite remarkable, the main insight of this chapter is that it is often possible to combine the predictions of the models in a way that achieves smaller code length even than $\hat{k}$! This can be done if the index of the best predicting model *changes with the sample size n in a predictable way*. Such cases are common in model selection, for example with nested models. If $\mathcal{M}_1 \subset \mathcal{M}_2$ then p_1 may predict better than p_2 at small sample sizes (roughly because $\mathcal{M}_2$ has more parameters that need to be estimated than $\mathcal{M}_1$), while p_2 may give better

predictions at large sample sizes (because $\mathcal{M}_2$ can fit more patterns in the data). This phenomenon is essentially just the *bias-variance trade-off*. The behaviour of Bayesian model averaging in such a setting is illustrated by Figure 2.1.

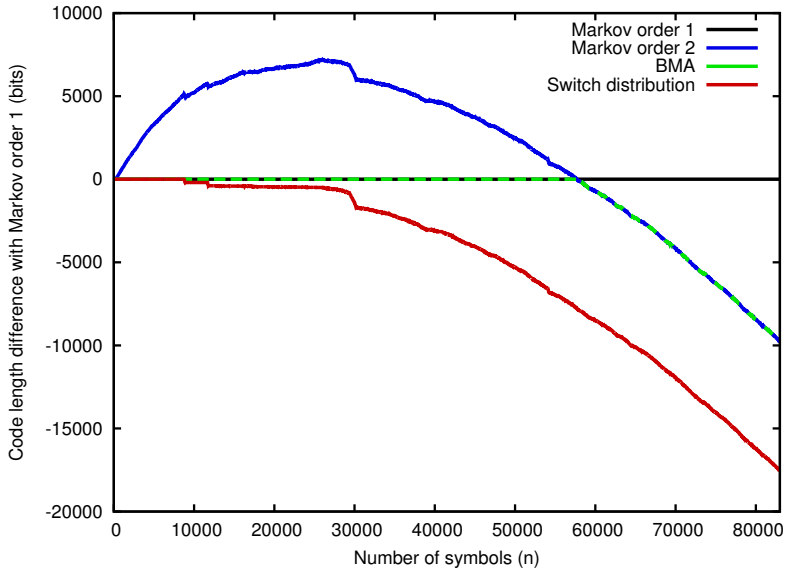


Figure 2.1: The Catch-up Phenomenon

The figure compares the code lengths for two Markov chain models of different order on the first n characters of Lewis Carroll's "Alice's Adventures in Wonderland" as a function of n, where each character in the book is considered an outcome[1]. It shows the code length difference $-\log p_2(x^n) - (-\log p_1(x^n))$, where p_k is the Bayesian marginal likelihood for the model $\mathcal{M}_k$ containing the k-th order Markov chains, parametrised by their transition probabilities. The book uses 84 distinct symbols. For simplicity we used uniform (Dirichlet$(1, 1, \ldots, 1)$) priors, but the same phenomenon occurs for other common priors such as Jeffreys' prior. The graph is restricted to the first half of the book only, to highlight the region of interest; the full text is 166 926 characters long.[2]

[1] An e-book version of "Alice's Adventures in Wonderland" was made available by project Gutenberg at www.gutenberg.org.

[2] The total code lengths for the full book are 603 906 and 554 494 bits for the first and

Note that if the difference in code length increases over an interval, this means that on average p_1 is making better predictions of those outcomes than p_2, and vice versa. To select the best predictor, one would therefore like to estimate the (sign of the) *derivative* of the graph. We see that on the first 26 000 outcomes, p_1 gets ahead by about 7 200 bits, but that p_2 predicts better afterwards. Ideally, we would therefore like to predict the first 26 000 outcomes like p_1 and then *switch* to predicting like p_2 for the remainder of the novel. However, BMA (with prior $\pi(1) = \pi(2) = 1/2$ on the models) only starts to behave like p_2 when p_2 *catches up* with p_1 around $n = 58\,000$. This is explained by the fact that the posterior depends, not on the derivative, but on the *height* of the graph, and is exponentially concentrated on the model with smallest code length. The result is that, between the maximum of the graph and the point where it reaches zero, p_{bma} behaves like p_1 while p_2 is making better predictions: since at $n = 26\,000$, p_2 is 7 200 bits behind, and at $n = 58\,000$, it has caught up, in between p_2 must have outperformed p_1 by 7 200 bits!

Note that the models $\mathcal{M}_1$ and $\mathcal{M}_2$ in this example are very crude; for this particular application much better models are available. Figure 2.1 is intended as a simple illustration of the catch-up phenomenon only. However, the general phenomenon that different models predict better at different sample sizes occurs widely, both in theoretical settings and on real-world data. For example, we have encountered the same catch-up phenomenon in regression with polynomials (see Section 2.7), and in unreported experiments to select the number of bins in histogram density estimation. We argue that failure to take this effect into account explains the suboptimal convergence rates of Bayes factors model selection and related methods. In Section 2.2 we define an alternative way of combining two distributions p_1 and p_2 into a single distribution p_{sw}, which we call the *switch distribution*. Figure 2.1 shows that the switch distribution first predicts roughly like p_1, but switches to p_2 almost immediately after it starts making better predictions.[3] It essen-

second order Markov chains, respectively, and 554 495 and 546 698 bits for BMA and the switch distribution.

[3] In fact, p_2 already slightly outperforms p_1 over short sequences of outcomes before $n = 26\,000$. This is exploited by the switch distribution, which can switch back and forth between the available predictors if necessary (see Section 2.2.2). The sharp drop around sample size 29 100 corresponds to "The Mouse's Tale" which uses long strings of spaces for unusual indentation, a structure that cannot be represented well by a first

tially does this *no matter what sequence x^n is actually observed*. The switch distribution is a modification of the Bayesian marginal distribution that assigns positive prior weight to predicting with different models at different sample sizes, instead of putting all prior weight on prediction with the same model for all sample sizes, like BMA. This allows us to avoid the implicit, and often wrong, a priori assumption that a single model will be the best predictor at all sample sizes. After conditioning on data, the posterior we obtain therefore gives a better indication of which model predicts best *at the actual sample size*, and hence achieves smaller risk. Indeed, the switch distribution, when viewed in terms of the sequential predictions it induces, is closely related to earlier algorithms for *tracking the best expert* in the universal prediction literature [Koolen and de Rooij, 2008a, Herbster and Warmuth, 1998, Vovk, 1999, Volf and Willems, 1998, Cesa-Bianchi and Lugosi, 2006]; however, both the context in which we apply the switch distribution and the theorems that we prove, are very different.

2.1.3 Overview

In Section 2.2 we define the switch distribution and give an explicit algorithm for its practical application. While we switched between only two models in the example above, the general definition allows switching between any countable number of models. The predictions for each model may either be based on the Bayesian predictive distribution or on parameter estimation, like for example maximum likelihood. This is explained in Section 2.3, which also discusses model selection in the sequential prediction setting. A first (minor) result is presented in Section 2.4, were we define minimax (cumulative) risk and it is shown that, like Bayesian model averaging, the switch distribution achieves the minimax cumulative risk in typical parametric settings. Our main cumulative risk convergence results, however, are for nonparametric model classes. These results, which are presented in Section 2.5, apply regardless of whether prediction is based on the Bayesian predictive distribution or on parameter estimation. They are followed by our main consistency result in Section 2.6, which only applies to Bayesian prediction strategies. Section 2.7 contains a simulation study of linear regression with polynomials. The discussion in Section 2.8 puts our

order Markov chain.

work in a broader context and explains how it fits into the existing literature. In particular, Section 2.8.3 shows how the switch distribution may be further modified to achieve the minimax *instantaneous* rather than cumulative risk. We end with a brief conclusion. The proofs of all results are at the end of the chapter, in Sections 2.9 and 2.10.

2.2 The Switch Distribution

2.2.1 Preliminaries

For any set $\mathcal{S}$, let $\mathcal{S}^n$ denote the n-fold Cartesian product, let $\mathcal{S}^* := \bigcup_{n=0}^{\infty} \mathcal{S}^n$ and let $\mathcal{S}^\infty$ denote the (uncountable) set of infinite sequences over $\mathcal{S}$. Analogously, let x^n denote an n-tuple $x_1, \ldots, x_n$ (x^0 is the empty sequence) and let x^∞ denote an infinite sequence.

Consider a random process $X^\infty \in \mathcal{X}^\infty$, where each outcome takes values in a space $\mathcal{X} \subseteq \mathbb{R}^d$ of finite dimension $d \in \mathbb{Z}^+ = \{1, 2, \ldots\}$. We call p a (sequential) *prediction strategy* for X^∞ if it issues a density $p(x_{n+1} \mid x^n)$ on $x_{n+1} \in \mathcal{X}$ for all $x^n \in \mathcal{X}^*$. If the data are assumed to be drawn from a distribution p^* we sometimes call the prediction strategy p an *estimator* to emphasize that p is intended to approximate p^*. For simplicity, we assume throughout that this density is taken relative to either the usual Lebesgue measure (if $\mathcal{X}$ is continuous) or the counting measure (if $\mathcal{X}$ is countable). In the latter case $p(x_{n+1} \mid x^n)$ is a probability mass function. Such sequential prediction strategies are sometimes called prequential forecasting systems [Dawid, 1984]. An instance is given in Example 2.2 below.

Our notation emphasises that the conditional densities of a distribution may always be viewed as a prediction strategy; vice versa, the predictions of any prediction strategy p may be viewed as the conditional probabilities of a distribution for X^∞ with density

$$p(x^n) = p(x_1) \cdot p(x_2 \mid x_1) \cdot \ldots \cdot p(x_n \mid x^{n-1}). \tag{2.4}$$

With some abuse of notation, we also use the symbol p to denote this distribution. For countable sample spaces, such a distribution can always be defined; for uncountable $\mathcal{X}$ we require the following standard measurability assumption: for any $n \in \mathbb{Z}^+$ and any fixed measurable event $A_{n+1} \subseteq \mathcal{X}$ the probability $p(A_{n+1} \mid x^n)$ should be a measurable function of x^n (see e.g. [Shiryaev, 1996, p. 249, Theorem 2]).

2.2.2 Definition

We start with a given, countable set of prediction strategies $\{p_k \mid k \in \mathcal{A}\}$; see Example 2.1 below for a concrete case. Based on the set $\{p_k \mid k \in \mathcal{A}\}$, we first define a new family $\mathcal{Q} = \{q_{\mathbf{s}} \mid \mathbf{s} \in \mathbb{S}\}$ of prediction strategies that switch between them. The parameter set $\mathbb{S}$ for these switching strategies is defined as

$$\mathbb{S} = \Big\{((t_1,k_1),\ldots,(t_m,k_m)) \in (\mathbb{Z}^+ \times \mathcal{A})^m \mid m \in \mathbb{Z}^+, 1 = t_1 < \ldots < t_m\Big\}. \tag{2.5}$$

Each parameter $\mathbf{s} \in \mathbb{S}$ specifies the indices $k_1,\ldots,k_m$ of m original prediction strategies to be used by $q_{\mathbf{s}}$ in sequence, and the sample sizes $t_1,\ldots,t_m$ at which switches occur from one strategy to the next. Formally,

$$q_{\mathbf{s}}(x_{n+1} \mid x^n) = p_{k_j}(x_{n+1} \mid x^n) \qquad \begin{array}{l}\text{for the largest } j \le m \\ \text{such that } t_j \le n+1.\end{array} \tag{2.6}$$

For example, t_4 is the index of the first outcome that is predicted using p_{k_4}. The extra switch-point t_1 is included to simplify boundary cases; we fix $t_1 = 1$ so that k_1 represents the strategy that is used first, before any actual switch takes place. Thus the total number of switches is $m-1$. Switching to the same predictor multiple times (consecutively or not) is allowed.[4]

The switch distribution is a Bayesian mixture of the elements of $\mathcal{Q}$ according to a prior π on $\mathbb{S}$:

Definition 2.1 (Switch Distribution)**.** The *switch distribution* p_{sw}, defined with respect to a prior probability mass function π on $\mathbf{s}$, is the distribution for $(X^\infty, \mathbf{s})$ with density

$$p_{\mathrm{sw}}(x^n, \mathbf{s}) := q_{\mathbf{s}}(x^n)\,\pi(\mathbf{s}) \tag{2.7}$$

for any $x^n \in \mathcal{X}^*$, and $\mathbf{s} \in \mathbb{S}$.

Hence the marginal switch distribution on n outcomes has density

$$p_{\mathrm{sw}}(x^n) = \sum_{\mathbf{s}\in\mathbb{S}} q_{\mathbf{s}}(x^n)\,\pi(\mathbf{s}). \tag{2.8}$$

[4] It is not necessary here, but the definitions and the algorithm can be modified to disallow such reflexive switches.

By Bayes' theorem, the prior π, conditioned on observed data x^n, induces a posterior distribution $p_{\mathrm{sw}}(\mathbf{s} \mid x^n) \propto q_{\mathbf{s}}(x^n)\pi(\mathbf{s})$ on switching strategies $\mathbf{s}$. The marginal of this posterior on the prediction strategy that is used to predict the next outcome will be of special interest. For $\mathbf{s} = \Big((t_1, k_1), \ldots, (t_m, k_m)\Big)$, define the random variable $K_n(\mathbf{s}) = k_{j^*}$, where j^* is the largest j such that $t_j \leq n$. Thus, $K_n(\mathbf{s})$ is the prediction strategy that is used by $q_{\mathbf{s}}$ to predict the n-th outcome. We can then consider, say, the posterior probability assigned to each prediction strategy upon observing x^n:

$$p_{\mathrm{sw}}(K_{n+1} = k \mid x^n) = \left(\sum_{\mathbf{s}: K_{n+1}(\mathbf{s}) = k} p_{\mathrm{sw}}(x^n, \mathbf{s}) \right) / p_{\mathrm{sw}}(x^n). \tag{2.9}$$

This quantity is computed by the algorithm presented in Section 2.2.5; it is also used to define a model selection criterion based on the switch distribution in Section 2.3.

2.2.3 Structure of the Prior

Partly to allow for an efficient algorithm (see Section 2.2.5), and partly because it facilitates our further results, we require that π can be written in the form

$$\begin{aligned} &\pi((t_1, k_1), \ldots, (t_m, k_m)) \\ &\qquad = \mu(m) \left(\prod_{j=1}^{m-1} \kappa_{t_j}(k_j) \tau(Z = t_{j+1} \mid Z > t_j) \right) \lambda_{t_m}(k_m). \end{aligned} \tag{2.10}$$

Here, μ is a prior probability mass function on the number of prediction strategies m, which is equal to the number of switches plus one. Further, τ is a prior mass function on the switching indices, which are the integers greater than one, and for all $n \in \mathbb{Z}^+$, κ_n is a prior mass function on some subset of strategies indexed by $\mathcal{K} \subseteq \mathcal{A}$ and λ_n is a prior on some subset of strategies indexed by $\mathcal{L} \subseteq \mathcal{A}$. The set $\mathcal{K}$ indexes the prediction strategies that can be switched to while switching has not yet stabilized, i.e. if one will switch at least once more in the future. The set $\mathcal{L}$ indexes the set of *final* prediction strategies that can

be switched to at the last switch. We sometimes blur the distinction between prediction strategies and their indices and say, for example, that $\mathcal{K}$ "contains" prediction strategies.

In the *basic* version of the switch distribution, we do not distinguish between $\mathcal{L}$ and $\mathcal{K}$, and set $\mathcal{L} = \mathcal{K} = \mathcal{A}$. For our convergence rate results, however, we will consider advanced versions of the switch distribution, in which $\mathcal{L}$ is still a given set of prediction strategies, but $\mathcal{K}$ contains slightly modified versions of the prediction strategies in $\mathcal{L}$. These will be introduced in Section 2.5.1. It will then become necessary to allow κ_n to depend on n. For computational reasons it may also be convenient to allow λ_n to depend on n (since no computation is necessary for prediction strategies with zero prior probability), and we therefore allow this in our definitions and theorems. All our results, however, are easiest to understand when λ_n does not depend on n.

Our algorithm, and consistency and convergence rate theorems all impose further conditions on the prior π, which will be stated in each case. For concreteness, we remark that every prior of the following form is compatible with all our results in the following sections:

$$\mu(m) = 2^{-m}, \qquad \tau(n) = \frac{1}{n(n-1)}, \qquad \text{and } \kappa_n \text{ and } \lambda_n \text{ are uniform on their support,} \tag{2.11}$$

as long as the supports of κ_n and λ_n never shrink with n and are at most of polynomial size in n.

Example 2.1. In the Markov chain example of Figure 2.1, p_{sw} is instantiated as follows. We set $\mathcal{L} = \mathcal{K} = \mathcal{A} = \{1, 2\}$, and define the prior π using (2.11), where the support of κ_n and λ_n is equal to $\mathcal{A}$ for all n. For $k \in \mathcal{A}$, p_k, as used in (2.6), is defined as the Bayesian marginal likelihood (see (2.1)) relative to the k-th order Markov model equipped with the uniform prior. The p_k are viewed as prediction strategies by defining $p_k(x_{n+1} \mid x^n) = p_k(x^{n+1})/p_k(x^n)$, such that the corresponding distribution is the standard *Bayesian predictive distribution* after conditioning on observations x^n [Bernardo and Smith, 1994].

In all applications in this chapter, the prediction strategies p_k will be based on (parametric) models $\mathcal{M}_k$. They will either be Bayesian predictive distributions as in Example 2.1, or parameter estimators relative to $\mathcal{M}_k$, as explained in Section 2.3. Note however that, in principle, the

switch distribution may be applied to completely arbitrary prediction strategies: p_k could just as well represent the prediction of next day's probability of rain as issued by a weather forecaster on television.

2.2.4 Comparison to Bayesian model averaging

As discussed in the introduction, one advantage of averaging over a set of predictors $\mathcal{P} = \{p_1, p_2, \ldots\}$ using p_{bma} is that it guarantees a bound $-\log \pi(\hat{k})$ on the difference in code length with the best predictor $p_{\hat{k}}$. This property is shared by p_{sw}, which multiplicatively dominates p_{bma}. To see this, let $\mathcal{L} = \mathcal{P}$ and define λ_1 to be equal to the prior used in p_{bma}. (The set $\mathcal{K}$ may be arbitrary, for example equal to $\mathcal{L}$.) Then comparison with the switch distribution shows that BMA corresponds to using a prior that allows no switches at all between predictors. This corresponds to the case $m = 1$ in the prior from (2.10). We therefore find that

$$\begin{aligned} p_{\text{sw}}(x^n) &\geq \sum_{\mathbf{s} \in \{((1,k)) | k \in \mathcal{L}\}} \pi(\mathbf{s}) q_{\mathbf{s}}(x^n) \\ &= \mu(1) \sum_{k \in \mathcal{L}} \lambda_1(k) p_k(x^n) = \mu(1) p_{\text{bma}}(x^n) \end{aligned}$$

for all n, x^n. Thus, p_{sw} can be smaller than p_{bma} by at most a constant factor $\mu(1)$, which is the prior probability of never switching between predictors. The converse of this is not true however: as Figure 2.1 illustrates, the switch distribution may achieve substantially smaller code length than p_{bma}. This is also seen in the simulation study in Section 2.7.

2.2.5 Hidden Markov Model and Efficient Computation

The following material is mostly of practical interest and can be skipped by readers who wish to reach the more theoretical material with as few distractions as possible.

Under certain conditions on the prior π, the switch distribution can be represented by a *hidden Markov model* (HMM) [Rabiner, 1989] with state transition diagram as in Figure 2.2. This is the case if μ is geometric, i.e. $\mu(m) = \theta^{m-1}(1-\theta)$ for some $0 \leq \theta \leq 1$. For each time step, the hidden state of the HMM represents the following information: (1) the prediction strategy that is used to predict the outcome at that time step, and (2) whether or not further switches are still possible.

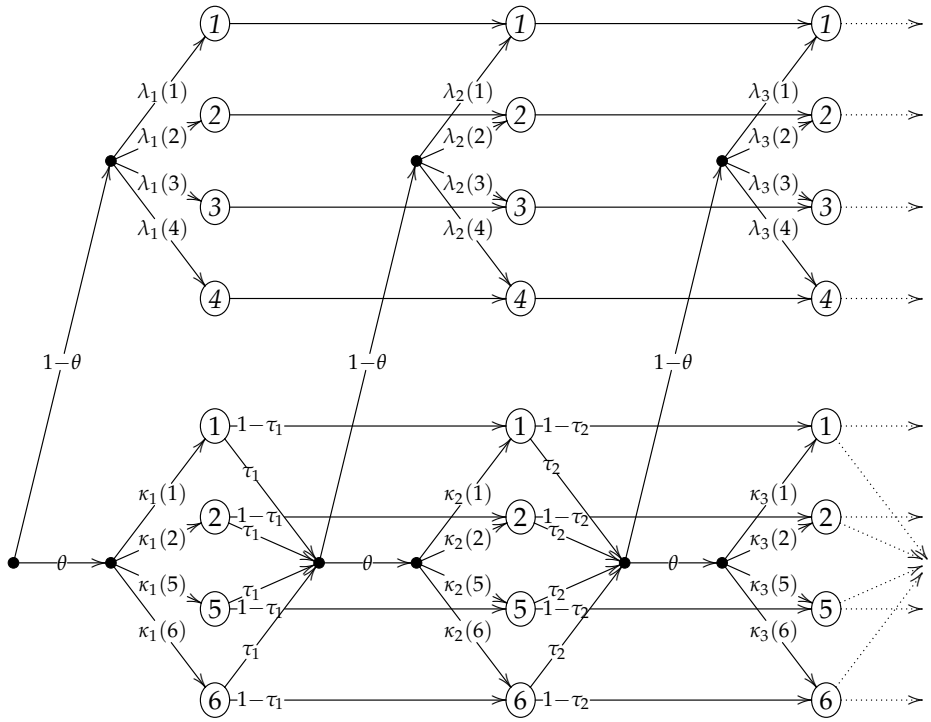


Figure 2.2: State transitions in the HMM for six prediction strategies

In Figure 2.2 there are six prediction strategies $\mathcal{A} = \{1,2,3,4,5,6\}$, and $\text{support}(\lambda_n) = \mathcal{L} = \{1,2,3,4\}$, $\text{support}(\kappa_n) = \mathcal{K} = \{1,2,5,6\}$ for all n; each column of numbered circles denotes the alternative values that the hidden states can take, italics indicate that no further switches are possible; and we abbreviate $\tau_i = \tau(Z = i+1 \mid Z > i)$. Note that the lower transitions in Figure 2.2 keep track of prediction strategies in $\mathcal{K}$, for which further switches will occur, whereas the upper part tracks prediction strategies in $\mathcal{L}$, for which switching has stabilized. See [Koolen and de Rooij, 2008b,a] for more details about this interpretation and a proof that the definition of the switch distribution used in this chapter and its definition in terms of an HMM coincide.

Most densities of interest, such as $p_{\text{sw}}(x_{n+1} \mid x^n)$, $p_{\text{sw}}(x^n)$ and $p_{\text{sw}}(K_{n+1} = k \mid x^n)$, are easy to obtain if we can sequentially compute the marginal density $p_{\text{sw}}(K_{n+1} = k, X^n = x^n)$ for all $n = 1, \ldots, N$, which can be done using the Forward Algorithm for HMMs [Rabiner, 1989].

Its instantiation for the switch distribution is given by Algorithm 2.1. The algorithm maintains weights for any prediction strategy that was assigned positive weight in the past. Let $\mathcal{S}_n^\lambda := \bigcup_{i=1}^n \text{support}(\lambda_n)$ and $\mathcal{S}_n^\kappa := \bigcup_{i=1}^n \text{support}(\kappa_n)$. The algorithm then runs as follows.

Algorithm 2.1 SWITCH(x^N)

1 **for** $k \in \mathcal{S}_1^\kappa$ **do** $w_\kappa[k] \leftarrow \kappa_1(k) \cdot \theta$ **end for**
2 **for** $k \in \mathcal{S}_1^\lambda$ **do** $w_\lambda[k] \leftarrow \lambda_1(k) \cdot (1-\theta)$ **end for**
3 **for** $n = 1, \ldots, N$ **do**
4 **for** $k \in \mathcal{S}_n^\kappa \cup \mathcal{S}_n^\lambda$ **do**
5 $v_\kappa \leftarrow w_\kappa[k]$ if $k \in \mathcal{S}_n^\kappa$, and 0 otherwise
6 $v_\lambda \leftarrow w_\lambda[k]$ if $k \in \mathcal{S}_n^\lambda$, and 0 otherwise
7 Output $(n, k, v_\kappa + v_\lambda)$ ▷ *Report* $p_{\text{sw}}(K_n = k, X^{n-1} = x^{n-1})$
8 **end for**
9 **for** $k \in \mathcal{S}_n^\kappa$ **do** $w_\kappa[k] \leftarrow w_\kappa[k] \cdot p_k(x_n \mid x^{n-1})$ **end for**
10 **for** $k \in \mathcal{S}_n^\lambda$ **do** $w_\lambda[k] \leftarrow w_\lambda[k] \cdot p_k(x_n \mid x^{n-1})$ **end for**
11 `pool` $\leftarrow \tau_n \cdot \sum_{k \in \text{support}(\kappa_n)} w_\kappa[k]$
12 **for** $k \in \mathcal{S}_{n+1}^\kappa \cup \mathcal{S}_{n+1}^\lambda$ **do**
13 $v_\kappa \leftarrow w_\kappa[k]$ if $k \in \mathcal{S}_n^\kappa$, and 0 otherwise
14 $v_\lambda \leftarrow w_\lambda[k]$ if $k \in \mathcal{S}_n^\lambda$, and 0 otherwise
15 $w_\kappa[k] \leftarrow v_\kappa \cdot (1-\tau_n) \; +$ `pool` $\cdot \kappa_n(k) \cdot \theta$
16 $w_\lambda[k] \leftarrow v_\lambda +$ `pool` $\cdot \lambda_n(k) \cdot (1-\theta)$
17 **end for**
18 **end for**
19 Compute and output $p_{\text{sw}}(K_{N+1} = k, X^N = x^N)$ as in lines 4–8.

The total running time of Algorithm 2.1 is $O(\sum_{n=1}^N |\mathcal{S}_n^\kappa| + |\mathcal{S}_n^\lambda|)$, which is linear in the number of outcomes N and the sizes of the supports. For example, if $\text{support}(\kappa_n) = \text{support}(\lambda_n) = \mathcal{A}$, then the running time is $|\mathcal{A}| \cdot O(N)$, which is typically of the same order as that of model selection criteria like AIC and BIC. For an example where the supports do depend on n, see Section 2.5.3, Example 2.4.

The algorithm may also be understood as one of a variety of *expert tracking algorithms* [Koolen and de Rooij, 2008b,a]. In fact, it may be viewed as a generalisation of the FIXED-SHARE algorithm [Herbster and Warmuth, 1998]: the main difference is that FIXED-SHARE does not include states in the HMM from which no further switches are possible (i.e. it fixes $\theta = 1$). However, a distinction between $\mathcal{K}$ and $\mathcal{L}$ is necessary to get a consistent method. In addition, whereas FIXED-SHARE always

uses a geometric prior τ with a parameter that needs to be tuned, we allow any choice of τ, which allows us to get a parameterless algorithm that achieves the minimax cumulative risk.

2.3 Model Selection, Prediction and Estimation

We consider a two-stage approach to inference based on a sequence of models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ In the first stage, for all $k = 1, 2, \ldots$, a single "meta" prediction strategy p_k is associated with each model $\mathcal{M}_k$. In the second stage, these prediction strategies are either used to select a single model based on the observed data x^n, or they are combined further into a "meta meta" prediction strategy for prediction of future outcomes. We treat these stages as orthogonal to gain flexibility, even though many methods described in the literature define both stages in tandem.

2.3.1 Stage 1: Models and Associated Prediction Strategies

We define a *model* $\mathcal{M}$ as a set of prediction strategies. A model is more commonly viewed as a set of distributions, but since distributions can be viewed as prediction strategies as explained above, we may think of a model as a set of prediction strategies as well. With each model, we associate a single "meta" prediction strategy; the models themselves are only used in terms of these meta strategies and are not referenced directly. Our results about predictive performance in Sections 2.4 and 2.5 apply regardless of how these meta strategies are defined; for our consistency result there are some restrictions that are explained in Section 2.6. We proceed with some important examples for parametric models $\mathcal{M} = \{p_\theta \mid \theta \in \Theta \subseteq \mathbb{R}^d\}$.

First, a natural approach is to define a *parameter estimator* $\hat{\theta} : \mathcal{X}^* \to \Theta$, which maps any data x^n of any length n to a "best guess" of the true/best parameter in the model. The next outcome is subsequently predicted using the strategy that is selected by the parameter estimator: $p(x_{n+1} \mid x^n) = p_{\hat{\theta}(x^n)}(x_{n+1} \mid x^n)$. Recall that by (2.4) this also defines a joint density $p(x^n) = p(x_1 \mid x^0) \cdot \ldots \cdot p(x_n \mid x^{n-1})$.

Second, the Bayesian approach to model selection or model averaging goes the other way around. Given a prior density w on Θ, it first

defines a joint density on x^n, called the *marginal likelihood*, as

$$p(x^n) = \int_{\theta \in \Theta} p_\theta(x^n) w(\theta) \, \mathrm{d}\theta. \tag{2.12}$$

This induces the Bayesian prediction strategy

$$p(x_{n+1} \mid x^n) = \frac{p(x^{n+1})}{p(x^n)} = \int_{\theta \in \Theta} p_\theta(x_{n+1} \mid x^n) w(\theta \mid x^n) \, \mathrm{d}\theta, \tag{2.13}$$

where $w(\theta \mid x^n) = p_\theta(x^n) w(\theta) / \int p_\theta(x^n) w(\theta) \, \mathrm{d}\theta$ is the posterior. If $p(x^n) = 0$, then the Bayesian prediction $p(x_{n+1} \mid x^n)$ is not defined. In practice this is usually of minor concern, either because $p(x^n)$ is positive for almost all x^n, or because one can make some reasonable default choice for $p(x_{n+1} \mid x^n)$ when it is not.

Example 2.2. Consider the Bernoulli model $\mathcal{M} = \{p_\theta \mid \theta \in [0,1]\}$ that regards $X_1, X_2, \ldots$ as a sequence of independent, identically distributed (i.i.d.) Bernoulli random variables taking values in $\mathcal{X} = \{0,1\}$, with $p_\theta(X_{n+1} = 1) = \theta$. Given past data x^n, we may predict x_{n+1} using the maximum likelihood (ML) estimator for x^n: $\hat{\theta}(x^n) = n^{-1} \sum_{i=1}^n x_i$, but then the prediction of x_1 is undefined, and the first outcome different from x_1 is assigned probability 0, yielding infinite code length. If we use a "smoothed" ML estimator, like the Laplace estimator $\hat{\theta}'(x^n) = (1 + \sum_{i=1}^n x_i)/(n+2)$, then all predictions become well defined and we are guaranteed finite loss. It is well-known that the prediction strategy $p_{\hat{\theta}'}$ equals the Bayesian predictive distribution based on a uniform prior. Thus in this special case prediction based on parameter estimation and the Bayesian prediction strategy coincide!

Using a model in terms of a single associated prediction strategy p is known as the *prequential approach to statistics* [Dawid, 1984] or *predictive MDL* [Rissanen, 1984]. Regardless of whether p is based on parameter estimation or on Bayesian predictions, we may usually think of it as a universal code relative to the model [Grünwald, 2007].

2.3.2 Stage 2: Model Based Prediction and Model Selection

Let $\mathcal{M}_1, \mathcal{M}_2, \ldots$ be parametric models, with associated prediction strategies $p_1, p_2, \ldots$ For example, $\mathcal{M}_k$ may be the set of all k-th order Markov chains, or it may be the set of k-bin histograms in a density

estimation setting, parametrised by the densities in the bins, or, in a regression setting, $\mathcal{M}_k$ may be the set of degree $(k-1)$ polynomials with standard normal noise. (In the regression setting the prediction strategies in the model are also given access to the explanatory variables; see Sections 2.5.5 and 2.7.) In general, the number of parameters in $\mathcal{M}_k$ does not need to be a straightforward function of k; for example, in a regression setting with two explanatory variables Z_1 and Z_2, we may first let $\mathcal{M}_{(k_1,k_2)}$ indicate the model of polynomials with terms of the form $\theta_{j_1,j_2} Z_1^{j_1} Z_2^{j_2}$ with $0 \leq j_1 \leq k_1, 0 \leq j_2 \leq k_2$, and then define $\mathcal{M}_k$ in terms of a suitable 1-to-1 correspondence between (k_1, k_2) and k.

Model based prediction means combining the "meta" prediction strategies $p_1, p_2, \ldots$ into yet another, "meta meta" prediction strategy p. Analogous to when the prediction strategies in the model were combined into a single prediction strategy associated with the model, we describe the two main methods to achieve this.

Model Selection Criteria Define a function $\delta : \mathcal{X}^* \to \mathcal{A}$ which maps any data x^n of any length n to a "best guess" of the true/best model. We can then predict the next outcome using the prediction strategy that is selected by δ: $p(x_{n+1} \mid x^n) = p_{\delta(x^n)}(x_{n+1} \mid x^n)$. This is the analogue of using a parameter estimator in stage 1; on this level we call such a function a *model selection criterion*. AIC, BIC and LOO are examples of model selection criteria; in a Bayesian setting reporting the full posterior distribution on the model index is usually advocated, but when pressed for a single answer, a Bayesian may report the "maximum a posteriori" (MAP) model (as in Bayes factors model selection), which is also a model selection criterion in the sense considered here.

If we assume that the data are sampled i.i.d. from some distribution p^*, then, in light of the discussion in Section 2.3.1, we may also think of the prediction strategy $p_{\delta(x^n)}(x_{n+1} \mid x^n)$ as a density estimator of p^*. As we will see in Section 2.5, the resulting estimators can be adaptive in a very strong sense.

Thus, model selection criteria can be used to define a prediction strategy, or as adaptive estimators. Yet they are also important in their own right as tools to determine, given a fixed data set, which model best explains these data; if that is the goal, an important property for a model selection criterion to have is *consistency*, which means that given enough data it always selects the true model, if there is one. (See also

Section 2.6.)

Model Averaging The strategies associated with the models can also be combined by taking a weighted mixture of their predictions. The prototypical example is Bayesian model averaging, in which the predictions associated with the models are weighted by the posterior probability of the model, as in (2.2). It has been found that prediction using model averaging often performs substantially better than prediction based on model selection (see, for example [Kontkanen et al., 2000]); for this reason, while strictly AIC is a model selection criterion, its definition is sometimes extended to assign weights to the models when it is used for prediction [Akaike, 1979] (see also Section 2.7).

2.3.3 Model Selection and Prediction with the Switch Distribution

Model selection and prediction with the switch distribution is very similar to normal Bayes factors model selection and Bayesian model averaging. There are two important differences: first, the posterior distribution is on the switch parameters S rather than simply on the models. After observing x^n, in order to obtain the weights of the prediction strategies to predict x_{n+1}, the posterior is marginalised using the random variable K_{n+1} as in (2.9). (Ignoring normalisation, it is these marginalised weights that Algorithm 2.1 keeps track of.)

A second difference is that the switch distribution can be defined with respect to more prediction strategies than just those corresponding to the models: in our results, the set $\mathcal{L}$ indexes the models, but the set $\mathcal{K}$ indexes a set of *variations* of the corresponding prediction strategies (see Section 2.5). Hence we define the following model selection criterion for the switch distribution, which selects a model index from $\mathcal{L}$ only:

$$\delta_{\mathrm{sw}}(x^n) = \arg\max_{k \in \mathcal{L}} p_{\mathrm{sw}}(K_{n+1} = k \mid x^n). \tag{2.14}$$

As mentioned, the goal for this model selection criterion is to select a model with index $k \in \mathcal{L}$ that is good specifically at predicting the next outcome x_{n+1}. In Section 2.6 we show that, under mild conditions, this model selection criterion is also consistent.

Prediction of x_{n+1} given x^n with the switch distribution is done using the predictive density $p_{\mathrm{sw}}(x_{n+1} \mid x^n) = p_{\mathrm{sw}}(x^{n+1}) / p_{\mathrm{sw}}(x^n)$. In

Section 2.5 we show that under mild conditions $p_{\mathrm{sw}}(x_{n+1} \mid x^n)$ asymptotically achieves the minimax cumulative Kullback-Leibler risk.

2.4 Risk Bounds: Preliminaries and Parametric Case

In this section we analyse the performance of the switch distribution in terms of cumulative Kullback-Leibler risk. We define the central notions of (parametric and nonparametric) model classes, Kullback-Leibler risk, and worst-case and minimax (cumulative) risk. We illustrate these by showing that, in the parametric case, like Bayesian model averaging, the switch distribution achieves the minimax cumulative risk under mild conditions. This serves as a preparation for Section 2.5, where we consider nonparametric model classes and show that *unlike* Bayesian model averaging, the switch distribution under mild conditions still achieves the minimax cumulative risk.

2.4.1 Model Classes

The setup is as follows. Suppose $\mathcal{M}_1, \mathcal{M}_2, \ldots$ is a sequence of parametric models with associated prediction strategies $p_1, p_2, \ldots$ as before. Let us write $\mathcal{M} = \cup_{k=1}^{\infty} \mathcal{M}_k$ for the union of the models. Although formally $\mathcal{M}$ is a set of prediction strategies, it will often be useful to consider the corresponding set of distributions for $X^\infty = (X_1, X_2, \ldots)$. With minor abuse of notation we will denote this set by $\mathcal{M}$ as well.

To test the predictions of the switch distribution, we will want to assume that X^∞ is distributed according to a distribution p^* that satisfies certain restrictions. These restrictions will always be formulated by assuming that $p^* \in \mathcal{M}^*$, where $\mathcal{M}^*$ is some restricted set of distributions for X^∞. (Note that in p^* and $\mathcal{M}^*$, the star is simply part of the name, not the Kleene star operator.)

For simplicity, we will also assume throughout that, for any n, the conditional distribution $p^*(X_n \mid X^{n-1})$ has a density (relative to the Lebesgue or counting measure) with probability one under p^*. For example, if $\mathcal{X} = [0, 1]$, then $\mathcal{M}^*$ might be the set of all product measures that have uniformly bounded densities with uniformly bounded first derivatives.

We call $\mathcal{M}$ and $\mathcal{M}^*$ *model classes*. In the *parametric* setting, we have $\mathcal{M}^* \subseteq \mathcal{M}$; we briefly consider this case in Example 2.3 and Sec-

tion 2.4.4. Our strongest risk convergence results however, presented in Section 2.5, deal with situations in which $\mathcal{M}^* \setminus \mathcal{M}$ is non-empty. We are mostly interested in cases where $\mathcal{M}^*$ represents what is commonly called a *nonparametric model class*. For a concrete example, see Section 2.5.5.

2.4.2 Risk

For two distributions p and q, the Kullback-Leibler (KL) divergence from p to q is defined as

$$D(p\|q) = \mathbf{E}_{Y\sim p}\left[\log \frac{p(Y)}{q(Y)}\right].$$

KL divergence is never negative, and reaches zero if and only if $p = q$. Given $X^{n-1} = x^{n-1}$, we measure how well any estimator p predicts X_n in terms of the KL divergence $D(p^*(X_n \mid x^{n-1})\|p(X_n \mid x^{n-1}))$ [Barron, 1998]. Taking an expectation over X^{n-1} leads to the standard definition of the *risk* of estimator p at sample size n relative to KL divergence:

$$r(p^*, p, n) = \mathbf{E}_{X^{n-1}\sim p^*}\left[D\left(p^*(X_n \mid X^{n-1})\|p(X_n \mid X^{n-1})\right)\right]. \tag{2.15}$$

In a sequential prediction setting, it is natural to consider not only the standard KL risk, but also the *cumulative risk*

$$R(p^*, p, n) = \sum_{i=1}^{n} r(p^*, p, i).$$

The cumulative risk is equal to the information theoretic redundancy, i.e. the Kullback-Leibler divergence on n outcomes (see e.g. [Barron, 1998] or [Grünwald, 2007, Chapter 15]): for all n it holds that

$$\begin{aligned} R(p^*, p, n) &= \sum_{i=1}^{n} \mathbf{E}_{p^*}\left[\log \frac{p^*(X_i \mid X^{i-1})}{p(X_i \mid X^{i-1})}\right] \\ &= \mathbf{E}_{p^*}\left[\log \prod_{i=1}^{n} \frac{p^*(X_i \mid X^{i-1})}{p(X_i \mid X^{i-1})}\right] = D\left(p^{*(n)}\|p^{(n)}\right), \end{aligned} \tag{2.16}$$

where the superscript (n) indicates that $p^{*(n)}$ and $p^{(n)}$ are distributions for X^n. This implies the following proposition, which underlies all our convergence rate results:

Proposition 2.1. *Let p_1 and p_2 be densities on n outcomes. Suppose that p_1 dominates p_2 by a factor of $c \in (0,1]$, i.e. for all $x^n \in \mathcal{X}^n$, $p_1(x^n) \geq c \cdot p_2(x^n)$. Then for every p^*, $R(p^*, p_1, n) \leq R(p^*, p_2, n) - \log c$.*

Note that the proposition does not require the sequence X^∞ to be independent and identically distributed (i.i.d.) under p^*.

Example 2.3. As we observed in Section 2.2.4, the switch distribution dominates Bayesian model averaging by a factor $\mu(1)$. By Proposition 2.1 this means that, for all distributions p^*, irrespective of whether $p^* \in \mathcal{M}$ or not, $R(p^*, p_{\mathrm{sw}}, n) \leq R(p^*, p_{\mathrm{bma}}, n) - \log \mu(1)$. Thus the cumulative risk of the switch distribution is bounded by the risk of Bayes up to a constant that does not depend on n. In fact, our results in Section 2.5 imply that in nonparametric model averaging, the switch distribution achieves substantially *smaller* cumulative risk than p_{bma}. Furthermore, our experiments (Section 2.7) suggest that in practice also in the parametric case (i.e. $p^* \in \mathcal{M}_k$ for some k), the cumulative risk of the switch distribution may be substantially smaller than that of Bayesian model averaging, but we have no general theorems to substantiate this. A difficulty in formulating such a result is that, as we shall see in Section 2.4.4, for any parametric model $\mathcal{M}_k$, in the worst case over $p^* \in \mathcal{M}_k$, the cumulative risks of p_{sw} and p_{bma} are of comparable size.

2.4.3 Minimax Risk Convergence

We have defined the risk of an estimator p with respect to a fixed distribution p^*, but we are really interested in investigating the behaviour of the standard risk and the cumulative risk of the switch distribution in the worst case over all possible $p^* \in \mathcal{M}^*$. Define the worst case instantaneous risk and worst-case cumulative risk of an estimator p as, respectively,

$$r_{\mathrm{m}}(p, n) = \sup_{p^* \in \mathcal{M}^*} r(p^*, p, n); \qquad R_{\mathrm{m}}(p, n) = \sup_{p^* \in \mathcal{M}^*} \sum_{i=1}^{n} r(p^*, p, i).$$

Note that the supremum is taken outside of the sum: we consider worst-case cumulative risk rather than cumulative worst-case risk, which is unreasonably adversarial in the sequential setting. The corresponding minimax risk notions are obtained by minimising the worst-

case risk:

$$r_{\mathrm{mm}}(n) = \inf_p r_{\mathrm{m}}(p,n); \qquad R_{\mathrm{mm}}(n) = \inf_p R_{\mathrm{m}}(p,n),$$

where the infimum is over all possible estimators, as defined in Section 2.2.1. (Note that p is not required to be a member of $\mathcal{M}^*$ or $\mathcal{M}$.) Minimax cumulative risk has previously been studied by, among others, Haussler and Opper [1997], Rissanen et al. [1992], Barron [1998], Yang and Barron [1999] and Poland and Hutter [2005].

Our results below are interesting only if $R_{\mathrm{mm}}(n)$ is finite, which implies that $r_{\mathrm{mm}}(i) \leq R_{\mathrm{mm}}(n)$ should be finite as well, for all $i \leq n$. Conversely, finiteness of $r_{\mathrm{mm}}(1)$ implies finiteness of $r_{\mathrm{mm}}(i)$ for all $i \geq 1$ and hence finiteness of $R_{\mathrm{mm}}(n) \leq \sum_{i=1}^{n} r_{\mathrm{mm}}(i)$. Thus, in all results below, whenever we refer to a model class $\mathcal{M}^*$, we implicitly assume that $r_{\mathrm{mm}}(1)$ is finite.

To conveniently compare asymptotic behaviour of functions we use the following notation:

Definition 2.2. For two nonnegative functions $g, h : \mathbb{Z}^+ \to \mathbb{R} \cup \{\infty\}$, we write $g \preceq h$ or $h \succeq g$ if for all $\epsilon > 0$ there exists an n_0 such that $g(n) \leq (1+\epsilon)h(n)$ for all $n \geq n_0$.

Like ordinary inequality, $\preceq$ is reflexive ($f \preceq f$ for all f) and transitive ($f \preceq g$ and $g \preceq h$ implies $f \preceq h$). Note that $g \preceq h$ is equivalent to $\limsup_{n\to\infty} g(n)/h(n) \leq 1$ as long as $h(n)$ is never zero, and that $g \leq h$ implies $g \preceq h$.

We can now easily define the two notions of minimax risk convergence that are of interest in this chapter we say that an estimator p *achieves the minimax risk up to factor* c if $r_{\mathrm{m}}(p,n) \preceq c \cdot r_{\mathrm{mm}}(n)$, and similarly, p *achieves the minimax cumulative risk up to factor* c if $R_{\mathrm{m}}(p,n) \preceq c \cdot R_{\mathrm{mm}}(n)$. See Section 2.8.3 for further discussion of the relationships between these two convergence notions.

2.4.4 The Parametric Case

As shown by Clarke and Barron [1990], for a d-dimensional parametric family $\mathcal{M}_k = \{p_\theta \mid \theta \in \Theta_k\}$, under suitable regularity conditions, including a compactness condition on Θ_k, the cumulative risk of Bayesian prediction strategies p_k defined as in (2.12) satisfies, uniformly for all

$p^* \in \mathcal{M}_k$:

$$R(p^*, p_k, n) = \frac{d}{2} \log n + O(1),$$

for continuous prior densities $w(\theta) > 0$. (See also (1.24) in Chapter 1.) In [Clarke and Barron, 1994] they further show that the minimax cumulative risk relative to the model class $\mathcal{M}^* := \mathcal{M}_k$ satisfies $R_{\text{mm}}(n) = (d/2) \log n + O(1)$ as well. It follows that p_k achieves the minimax cumulative risk relative to $\mathcal{M}_k$. Since p_{bma} dominates p_k (by a factor determined by its prior probability), Proposition 2.1 implies that p_{bma} also achieves the minimax cumulative risk up to factor 1; and since (by Example 2.3) p_{sw} dominates p_{bma}, the switch distribution achieves the same.

In the nonparametric case, where p^* is in none of the considered models, the minimax optimal cumulative risk grows more quickly. Then the cumulative risk of p_{bma} may not be minimax optimal anymore [Rissanen et al., 1992] whereas, as we show in the following section, under mild conditions, the cumulative risk of p_{sw} is.

2.5 Two Cumulative Risk Bounds

In this section we present our risk bounds for nonparametric adaptive estimation based on the switch distribution. We first need to introduce the notion of "frozen" prediction strategies, that keep issuing the same prediction even as they are conditioned on more and more data, which will be required in the proofs of both cumulative risk theorems. We then introduce the notion of an oracle, which is essentially a model selection criterion augmented with knowledge of the true distribution. Theorem 2.1, our strongest cumulative risk result, is presented in Section 2.5.3. As mentioned in the introduction it requires augmenting the set of considered prediction strategies with linearly many frozen strategies, leading to a slower algorithm. A faster, but somewhat weaker, alternative is provided by Theorem 2.2 in Section 2.5.4.

2.5.1 Frozen Strategies

In the definition of the switch distribution we distinguished between $\mathcal{K}$, which indexes prediction strategies from which one will switch at least once more in the future, and the set $\mathcal{L}$, which indexes the set of

final prediction strategies that can be switched to at the last switch. In the basic version of the switch distribution, we set $\mathcal{L} = \mathcal{K}$. This version works well empirically, and can be proved to achieve the minimax cumulative risk in some particular nonparametric settings (such as those of Barron and Sheu [1991]; see [Van Erven et al., 2008a] for details). Yet it is hard to prove general results about its risk behaviour, for reasons we explain below. To make the switch distribution more amenable to mathematical analysis, we allow $\mathcal{K}$ to contain "frozen" (explained below) versions of the strategies in $\mathcal{L}$, so that $\mathcal{K} \neq \mathcal{L}$. Employing frozen strategies allows us to prove convergence rate results for quite general settings. Since our definition of frozen strategies only applies to i.i.d. data, we will restrict to this setting for the remainder of this section:

Definition 2.3 (Standard IID)**.** We call a *distribution* p^* for $X^\infty = X_1, X_2, \ldots$ "standard IID" if the random variables $X_1, X_2, \ldots$ are independent and identically distributed under p^*, and $p^*(X_1)$ has a density (relative to the Lebesgue or counting measure). We call a *model class* $\mathcal{M}^*$ "standard IID" if all $p^* \in \mathcal{M}^*$ are standard IID. For any two standard IID distributions p^*, p, we abbreviate $D(p^* \| p) := D(p^*(X_1) \| p(X_1))$.

For sufficiently regular i.i.d. models and suitable estimators p_k, the risk $r(p^*, p_k, n)$ converges to $\inf_{p \in \mathcal{M}_k} D(p^* \| p)$, the smallest risk obtainable by any distribution within $\mathcal{M}_k$. Roughly, the larger n, the more data available to base the prediction $p_k(x_{n+1} \mid x^n)$ on, and the smaller the risk $r(p^*, p_k, n)$. However, it turns out that the risk does not always decrease monotonically; for an example of temporarily increasing risk, see [Barron, 1998, Section 7]. The proof techniques we have developed, however, only apply if $r(p^*, p_k, n)$ is either nonincreasing or increases only very little in that $\sup_{k \in \mathcal{A}}(r(p^*, p_k, n+1) - r(p^*, p_k, n)) = O(1/n)$. To prove risk convergence rates, we could simply impose this condition on the predictors p_k, but, since it turns out to be hard to verify, this is not satisfactory. Instead, we therefore include modified prediction strategies whose risk can be guaranteed to be nonincreasing. This is achieved by "freezing" the issued predictions as follows.

Definition 2.4 (Frozen Strategies)**.** Let $\mathbf{t} = t_1, t_2, \ldots$ be a finite or infinite sequence of integers with $1 = t_1 < t_2 < t_3 < \ldots$ and let $|\mathbf{t}|$ denote the number of elements of the sequence; if this number is infinite, we have $|\mathbf{t}| = \infty$. Let $\{p_k \mid k \in \mathcal{L}\}$ be a set of prediction strategies. For each $k \in \mathcal{L}$, we define a new prediction strategy $p_{k \circ \mathbf{t}}$ by setting, for all

$x^{n+1} \in \mathcal{X}^*$, $p_{k\circ\mathbf{t}}(x_{n+1} \mid x^n) = p_k(x_{n+1} \mid x^{t_j-1})$, where $j \in \{1, \ldots, |\mathbf{t}|\}$ is the largest j' such that $t_{j'} \leq n+1$. We call $p_{k\circ\mathbf{t}}$ the "strategy p_k *frozen* at times $\mathbf{t}$".

Any reasonable estimator p_k based on parametric models $\mathcal{M}_k$ "learns" from experience, so that the predictions $p_k(x_{n+1} \mid x^n)$ depend on x^n; for the case that $\mathcal{M}_k$ is the Bernoulli model, this is illustrated in Example 2.2. If a strategy p_j is frozen at a single point in time t_0, i.e. $\mathbf{t} = t_0$, then the resulting strategy $p_{j\circ\mathbf{t}}$ "stops learning" at time t_0 and predicts using the same distributions for all $n \geq t_0$. If p_j is frozen at a sequence of time points $\mathbf{t} = t_1, t_2, \ldots$, then the resulting strategy $p_{j\circ\mathbf{t}}$ stops learning between t_1 and t_2, is brought up to date ('thawed') again at t_2, stops learning again between t_2 and t_3, and so on.

2.5.2 Oracles, Fast and Slow Switch Distribution

To apply our theorems below to a specific model class, one first has to define an *oracle* [Donoho and Johnstone, 1994] that achieves the desired cumulative risk. Model selection criteria are examples of oracles, but oracles are more powerful as they can additionally use knowledge about the true distribution p^*. In this chapter, we adopt a broad definition that gives the oracle full access to p^*:

Definition 2.5 (Oracle)**.** An oracle is a function $\omega : \mathcal{M}^* \times \mathcal{X}^* \to \mathcal{A}$ that, given not only the observed data $x^n \in \mathcal{X}^*$, but also the true distribution $p^* \in \mathcal{M}^*$, selects a prediction strategy $\omega(p^*, x^n)$. We say that ω is an oracle *relative to (prediction strategy sets)* $\mathcal{L}_1, \mathcal{L}_2, \ldots$ if for all $p^* \in \mathcal{M}^*$, all $n \geq 0$, all $x^n \in \mathcal{X}^n$, $\omega(p^*, x^n) \in \mathcal{L}_{n+1}$. We let $p_\omega(x_{n+1} \mid x^n) := p_{\omega(p^*,x^n)}(x_{n+1} \mid x^n)$ denote the prediction strategy associated with oracle ω.

We will compare the switch distribution to oracles relative to fixed strategy sets $\mathcal{L}_1 \subseteq \mathcal{L}_2 \subseteq \ldots \subseteq \mathcal{L}$. In the following two subsections, we establish minimax cumulative risk rates for two versions of the switch distribution. For both versions, we let $\mathcal{K}_n$ contain frozen versions of the prediction strategies in $\mathcal{L}_n$. In our theorems, we will define the switch distributions such that $\mathcal{K}_n$ and $\mathcal{L}_n$ are subsets of the supports of κ_n and λ_n, respectively, so that the switch distribution can mimic the behaviour of the oracle. In the *slow switch distribution*, we set

$$\mathcal{K}_n = \{k \circ t \mid k \in \mathcal{L}_n, t \in \{1, \ldots, n\}\}. \tag{2.17}$$

That is, for each $k \in \mathcal{L}_n$, the prediction strategies available at time n include versions of p_k frozen at all $t \leq n$. Note that for all $p_k \in \mathcal{L}_n$, the prediction strategy $p_{k \circ n} \in \mathcal{K}_n$ issues the same prediction for the next outcome. While in the basic switch distribution $\mathcal{K}_n = \mathcal{L}_n$, in the slow switch distribution $|\mathcal{K}_n| = n|\mathcal{L}_n|$. In processing this larger set of prediction strategies, the algorithm described in Section 2.2.5 becomes slower by a factor of $\Theta(n)$ compared to the basic switch distribution, which motivates the name "slow" — note that "slow" refers to running time rather than the rate at which switches take place. In Section 2.5.3 we show that, under weak conditions, the slow switch distribution achieves the minimax cumulative risk up to factor one, which is optimal.

In Section 2.5.4 we consider the *fast switch distribution*, in which $\mathcal{K}_n$ contains only a single frozen version of p_k for each $k \in \mathcal{L}_n$. The freezing times are chosen the same for all k and occur at exponentially increasing intervals. In this way we have $|\mathcal{K}_n| = |\mathcal{L}_n|$, which is the same as for the basic switch distribution. Thus the algorithm for the fast switch distribution is as fast as for the basic switch distribution. However, the faster running time (compared to the slow switch distribution) comes at a price: we can only prove that the fast switch distribution achieves the minimax cumulative risk under somewhat stronger conditions, and only up to a suboptimal constant factor.

For both the slow and fast switch distributions, we prove cumulative risk bounds below that depend on the following condition on the prior distribution used in the definition of p_{sw}:

Condition 2.1. The prior π of the switch distribution is defined as in (2.10) and satisfies

$$\begin{aligned} -\log \mu(m) &= O(m), \\ -\log \tau(t) &= O(\log t), \\ -\log \kappa_n(k) &= O(\log n) \quad \text{uniformly for all } k \in \mathcal{K}_n. \end{aligned}$$

This condition expresses that the tails of the distributions τ and κ_n are at least of polynomial thickness, and that the set of accessible prediction strategies $\mathcal{K}_n$ is at most polynomially large in n, which implies, if $\mathcal{K}_n$ and $\mathcal{L}_n$ are related as above, that $\mathcal{L}_n$ is also at most polynomially large. Thus, the number of models we can consider is at most polynomial in n.

Example 2.4. Suppose that $\mathcal{L}$ is countably infinite, e.g. $\mathcal{L} = \mathbb{Z}^+$. We may set, for example, $\mathcal{L}_n = \{1, \ldots, \lceil n^a \rceil\}$ for some finite $a > 0$. Note that the number of models of a given dimension may be large, as long as the total number of models equals $\lceil n^a \rceil$. Then, in order to satisfy Condition 2.1, we may make suitable choices for μ and τ and take $\lambda_n = \lambda$ and $\kappa_n = \kappa$ independent of n, for example as $\lambda(k) = 1/(k(k+1))$ and $\kappa(k \circ t) = 1/(k(k+1)\, t(t+1))$. Although this satisfies the condition, the predictions $p_{\mathrm{sw}}(x_{n+1} \mid x^n)$ cannot be computed by the algorithm of Section 2.2.5, which requires the supports of λ_n and κ_n to be finite. To apply the algorithm, we may instead reduce the supports of λ_n and κ_n to $\mathcal{L}_n$ and $\mathcal{K}_n$, respectively, and use the sample size dependent prior suggested in (2.11). The resulting running time for data $x_1, \ldots, x_n$ will then be of order $\sum_{i=1}^{n} (|\mathcal{K}_i| + |\mathcal{L}_i|)$; this is $O(n^{2+a})$ or $O(n^{1+a})$ for the slow and fast switch distributions, respectively.

2.5.3 Cumulative Risk Bound for Slow Switch Distribution

The cumulative risk of the slow switch distribution is asymptotically equal to that of any oracle, provided that the cumulative risk of that oracle is not too small:

Theorem 2.1 (Cumulative Risk for Slow Switch Distribution)**.** *Fix $\mathcal{L}_1 \subseteq \mathcal{L}_2 \subseteq \cdots \subseteq \mathcal{L}$ and define $\mathcal{K}_1, \mathcal{K}_2, \ldots$ as in (2.17). Let $\mathcal{M}^*$ be standard IID and suppose the switch distribution satisfies Condition 2.1. Then, for any oracle ω relative to prediction strategies $\mathcal{L}_1, \mathcal{L}_2, \ldots$ that satisfies*

$$\frac{(\log n)^{2+\alpha}}{R_{\mathrm{m}}(p_\omega, n)} \to 0 \tag{2.18}$$

for some $\alpha > 0$, the worst-case cumulative risk of the switch distribution grows no faster than the worst-case cumulative risk of ω:

$$R_{\mathrm{m}}(p_{\mathrm{sw}}, n) \preceq R_{\mathrm{m}}(p_\omega, n). \tag{2.19}$$

Note that every model selection criterion such as AIC or BIC that, at sample size n, is allowed to choose a model in $\mathcal{L}_n$, is a special case of an oracle relative to $\mathcal{L}_1, \mathcal{L}_2, \ldots$ Therefore, to make the theorem more concrete, it is useful to explicitly consider the case in which ω is in fact a model selection criterion. In that case, the condition (2.18) will be satisfied for all model classes $\mathcal{M}^*$ that are usually called "nonparametric":

for such model classes, the minimax risk $r_{\mathrm{mm}}(n)$ is typically of order $n^{-\alpha}(\log n)^{\beta}$ for some $0 < \alpha < 1$ and $\beta \in \mathbb{R}$ and thus satisfies $r_{\mathrm{mm}}(n) \succeq n^{-\gamma}$ for some $0 < \gamma < 1$. If ω is a model selection criterion, then (by Proposition 2.2 below) $R_{\mathrm{m}}(p_\omega, n) \geq R_{\mathrm{mm}}(n) \succeq n r_{\mathrm{mm}}(n) \succeq n^{1-\gamma}$, and (2.18) holds. Hence Theorem 2.1 implies the following:

Corollary 2.1. *Suppose $\mathcal{M}^*$ is a standard IID model class such that $(\log n)^{2+\alpha}/R_{\mathrm{mm}}(n) \to 0$ for some positive α, for example if $r_{\mathrm{mm}}(n) \succeq n^{-\gamma}$ for some $\gamma < 1$. Then for any model selection criterion $\delta : \mathcal{X}^{n-1} \to \mathcal{L}_n$, which selects only prediction strategies from $\mathcal{L}_n$, the worst-case cumulative risk of the switch distribution grows no faster than the worst-case cumulative risk of δ. That is,*

$$R_{\mathrm{m}}(p_{\mathrm{sw}}, n) \preceq R_{\mathrm{m}}(p_\delta, n), \tag{2.20}$$

where p_δ is the prediction strategy with predictions $p_{\delta(x^n)}(x_{n+1} \mid x^n)$.

In particular, for all model classes that are commonly called "non-parametric", the slow switch distribution performs at least as well as, for example, AIC and leave-one-out cross-validation (LOO). Note however that AIC and LOO always output a single model index whereas the switch distribution is allowed to predict using a weighted mixture of the p_k's. Let us consider in more detail an example where a Bayesian procedure with a sample size dependent prior achieves the minimax rate, and Theorem 2.1 implies that switching achieves the minimax rate as well, based on a prior that does not depend on the sample size.

To give but one example, Ghosal et al. [2008] analyse exponential families defined on $\mathcal{X} = [0,1]$. In their set-up, $\mathcal{M}_J$ is a log spline density model for splines of some fixed order q and resolution K, where $J = q + K - 1$, which is a $(J-1)$-dimensional exponential family. Now suppose that the true density p^* belongs to the class of α-smooth functions $C^\alpha[0,1]$. Ghosal et al. show that, at sample size n, a Bayes procedure with dimension $J_{n,\alpha} = \left\lfloor n^{1/(2\alpha+1)} \right\rfloor$ and, for each J, a fixed smooth prior w_J on the canonical parameters for $\mathcal{M}_J$, achieves the optimal rate of convergence $n^{-\alpha/(2\alpha+1)}$ in Hellinger distance. Since they make the further assumption that the density of p^* and all densities in $\mathcal{M}_J$ are uniformly bounded away from 0 and ∞, convergence in Hellinger risk at rate of order $r(n)$ implies convergence in instantaneous KL risk at rate of order $r(n)^2$ and vice versa [Barron and Cover, 1991]. Thus, they also achieve the optimal rate $n^{-2\alpha/(2\alpha+1)}$ in KL risk. Their procedure

is not "adaptive", since the prior depends on n and on the unknown smoothness α. Ghosal et al. [2008, page 75] show that, by putting a discrete prior μ_α on the set of rational-valued smoothnesses $\alpha \in \mathbb{Q}^+$, they can achieve the optimal rate up to a logarithmic factor. They write: "we believe that the logarithmic factor is not a defect of our proof, but connected to this prior... the logarithmic factor can be removed by using special *sample-size dependent* priors $\lambda_{n,\alpha}$ (depending on n) that put *less* mass on small models". Note that, if the belief of Ghosal et al. is correct, then sample-size independent priors also lead to an extra logarithmic factor in the cumulative KL rate of the Bayesian procedure. This may be viewed as an instance of the catch-up phenomenon. Indeed, if we use the switch distribution, we can achieve the cumulative minimax rate without extra logarithmic factor: we set $\mathcal{L} = \mathbb{Z}^+$, define the predictive distribution p_J based on the same prior w_J as Ghosal et al. and we use the prior $\lambda_n(J) = \lambda(J) = 1/J(J+1)$ and the corresponding $\kappa(J)$ as in the beginning of Example 2.4, and any suitable μ and τ such that Condition 2.1 holds. By Theorem 2.1, the switch distribution based on this prior adaptively achieves the optimal rate $n^{1/(2\alpha+1)}$ in the cumulative sense. Here we applied Theorem 2.1 with the oracle set to the procedure of Ghosal et al. with the sample size-dependent priors $\lambda_{n,\alpha}$, that are needed for Bayesian model averaging to achieve the minimax instantaneous rate; but the switch distribution itself avoids the use of any sample size-dependent priors to achieve that rate.

2.5.3.1 Remarks

1. Interestingly, the theorem and corollary also apply in the "misspecified" case in which $\mathcal{M}^*$ contains some p^* that cannot be approximated arbitrarily well by the list of models $\mathcal{M}_1, \mathcal{M}_2, \ldots$, i.e. if $\inf_{k \in \mathcal{L}, p \in \mathcal{M}_k} D(p^* \| p) > 0$. In that case, the cumulative risk of any oracle, including any model selection method, will increase, to first order, as αn for some $\alpha > 0$, and the cumulative risk of the switch distribution will increase as αn for the α achieved by the best oracle.

2. Condition 2.1 implies that $|\mathcal{K}_n| \leq n^a$ for some fixed $a > 0$. Since for the slow switch distribution $|\mathcal{L}_n| = |\mathcal{K}_n|/n$, this implies that the model selection criterion δ mentioned in Corollary 2.2 must output a model with index in a set with grows maximally polynomially in n. While it may grow superlinearly, it cannot grow exponentially,

which precludes application of the corollary in the general variable selection problem, where, at time n, one wants to select between a number of models that is exponential in n. This is discussed further in the Section 2.8.5.

3. The theorem is asymptotic, but by keeping track of constants one can also show that

$$R_{\mathrm{m}}(p_{\mathrm{sw}}, n) \leq 2R_{\mathrm{m}}(p_{\omega}, n) + c_1(\log n)^{2+\alpha} + c_2,$$

where the constants c_1 and c_2 depend on the prior. Thus, in this sense the cumulative risk of the switch distribution is close to that of the oracle for *every* n.

2.5.4 Cumulative Risk Bound for Fast Switch Distribution

To get minimax convergence rates for the fast switch distribution, we need to impose the following condition on the model class:

Condition 2.2. Relative to $\mathcal{M}^*$, the minimax risk r_{mm} does not decrease too fast in the sense that, for some nondecreasing, strictly positive function h_0 and constants $0 < c_1 \leq c_2$ and $0 \leq \gamma < 1$, it satisfies

$$c_1 h_0(n) \preceq n^{\gamma} r_{\mathrm{mm}}(n) \preceq c_2 h_0(n). \tag{2.21}$$

As can be seen by inspecting the proof of Theorem 2.2 below, this condition implies condition (2.18) and is therefore stronger. Yet, it is still weak enough to be satisfied by all model classes that are usually called nonparametric, including the regression setting discussed in Section 2.5.5 below. Note that it allows cases such as $r_{\mathrm{mm}}(n) = \Theta(n^{-\alpha}(\log n)^{\beta})$ for $\alpha < 1, \beta \in \mathbb{R}$. (For $\beta < 0$, take $\gamma > \alpha$ and let $h_0(n) = \Theta(n^{\gamma-\alpha}(\log n)^{\beta})$.) The smaller γ, the better the bound in Theorem 2.2 below.

If Condition 2.2 holds, we can establish minimax cumulative risk rates up to a constant factor c determined by the constants c_1 and c_2 and γ. The key here is the following relation between cumulative and instantaneous risk, proved in Section 2.9.3:

Proposition 2.2. *Suppose that $\mathcal{M}^*$ is a standard IID model class. Then*

$$r_{\mathrm{mm}}(n) \preceq n^{-1} R_{\mathrm{mm}}(n) \leq n^{-1} \sum_{i=1}^{n} r_{\mathrm{mm}}(i).$$

Furthermore, if $\mathcal{M}^*$ satisfies Condition 2.2 with constants c_1, c_2 and γ, and $r_{\mathrm{mm}}(n) < \infty$ for all n, then also

$$n^{-1} \sum_{i=1}^{n} r_{\mathrm{mm}}(i) \preceq \frac{c_2}{c_1} \frac{1}{1-\gamma} r_{\mathrm{mm}}(n).$$

Based on this proposition, in Section 2.9.4 we prove the following theorem:

Theorem 2.2 (Cumulative Risk for Fast Switch Distribution)**.** *Suppose $\mathcal{M}^*$ is a standard IID model class that satisfies Condition 2.2 with constants c_1, c_2 and γ. Suppose that there exists an oracle ω relative to sets of prediction strategies $\mathcal{L}_1 \subseteq \mathcal{L}_2 \subseteq \cdots \subseteq \mathcal{L}$ that achieves the minimax risk up to a non-decreasing function $f : \mathbb{Z}^+ \to [1, \infty)$, i.e. $r_{\mathrm{m}}(p_\omega, n) \preceq f(n)\, r_{\mathrm{mm}}(n)$, and is such that $r_{\mathrm{m}}(p_\omega, n) < \infty$ for all n. Let p_{sw} be the switch distribution with a prior that satisfies Condition 2.1, and with $\mathcal{K}_1, \mathcal{K}_2, \ldots$ defined as*

$$\mathcal{K}_n = \{k \circ \mathbf{t} \mid k \in \mathcal{L}_n\}$$

for an infinite increasing sequence $\mathbf{t} = t_1, t_2, \ldots$ with $t_1 = 1$ and $t_j \geq a \exp(bj)$ for positive constants a and b. Then the switch distribution achieves the minimax cumulative risk up to factor $c f(n)$ for a constant c. Specifically,

$$R_{\mathrm{m}}(p_{\mathrm{sw}}, n) \preceq c\, f(n)\, R_{\mathrm{mm}}(n),$$

with c given by

$$c = \left(\frac{c_2}{c_1}\right)^2 \cdot \frac{1}{1-\gamma} \sup_{j \geq 1} \left(\frac{t_{j+1} - 1}{t_j}\right)^{\gamma}. \tag{2.22}$$

In applications we can take, for example, $t_j = 2^{j-1}$, or, to get slightly better bounds, we may take $t_j = \max\{j, \lceil (1+\epsilon)^{j-1} \rceil\}$ for some small $\epsilon > 0$, so that the rightmost factor in (2.22) is bounded by $(1+\epsilon)^{\gamma}$. Analogously to Corollary 2.1, Theorem 2.2 implies the following:

Corollary 2.2. *Suppose $\mathcal{M}^*$ is a standard IID model class that satisfies Condition 2.2. Let the fast switch distribution be as in Theorem 2.2. If there exists any model selection criterion $\delta \colon \mathcal{X}^{n-1} \to \mathcal{L}_n$ at all that achieves the minimax risk up to a factor c_3, and δ has finite worst-case risk for all n, then the fast switch distribution achieves the minimax cumulative risk up to factor $c' = c \cdot c_3$, where c is as in (2.22), i.e.*

$$R_{\mathrm{m}}(p_{\mathrm{sw}}, n) \preceq c' R_{\mathrm{mm}}(n).$$

Thus, in typical nonparametric settings in which AIC or leave-one-out cross-validation achieve the minimax risk, the fast switch distribution also achieves this risk in the cumulative sense, albeit only up to a factor c', which may be larger than 1. Remarks analogous to remarks 1 through 3 below Corollary 2.1 apply to Corollary 2.2 as well.

2.5.5 Example: Gaussian Regression with Random Design

We now show that switching achieves the minimax cumulative risk in an important special case: Gaussian regression with random design relative to regression functions in certain Besov spaces. We do this by applying Theorems 2.1 and 2.2 to a result of Baraud [2002]. Readers who are not familiar with adaptive estimation theory may want to skip this section and focus on our practical regression experiments in Section 2.7 instead.

Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be independent, identically distributed pairs of random variables, with X_i taking values in $\mathcal{A} \subseteq \mathbb{R}^d$ for some $d > 0$ and Y_i real-valued. We assume that the explanatory variables X_i are all distributed according to a (possibly unknown) *design distribution* and that Y_i depends on X_i as

$$Y_i = f^*(X_i) + \xi_i,$$

where f^* is an unknown regression function from a set $\mathcal{F}^*$ of candidate functions, and the random variables $\xi_1, \xi_2, \ldots$ are distributed according to an *error distribution*, which we shall assume to be Gaussian with zero mean and known variance σ^2.

The set of candidate regression functions $\mathcal{F}^*$ is approximated by a finite or countably infinite number of linear spaces $\mathcal{F}_0, \mathcal{F}_1, \ldots$ of functions from $\mathcal{A}$ to $\mathbb{R}$. Here each $\mathcal{F}_k$ contains all linear combinations of the members of a finite orthonormal basis

$$S_k = \{\phi_{(k,1)}, \ldots, \phi_{(k,m_k)}\},$$

which contains m_k linearly independent functions from $\mathcal{A}$ to $\mathbb{R}$.

In one of his results, Baraud [2002] specializes this setting as follows: he lets $\mathcal{A} = [0, 1]$ and assumes that the design distribution has some density p^*_X that is bounded away from 0 and ∞. For some $r \geq 1$, he lets $\mathcal{F}_1, \mathcal{F}_2, \ldots$ each consist of piecewise polynomials of degree less than r on regular grids, where the grid width of functions in $\mathcal{F}_k$ is 2^{-k}.

In addition, we let $\mathcal{F}_0 = \{\mathbf{0}\}$ contain only the identically-0 function. Baraud defines Besov balls $\mathcal{B}_{\alpha,2,\infty}(R)$ of radius R in a certain Besov space that depends on α (see [Yang and Barron, 1999, DeVore and Lorentz, 1993] for the definition of Besov spaces). His Theorem 2.1 now implies that for this choice of $\mathcal{F}_0, \mathcal{F}_1, \ldots$, under certain assumptions, there exists an adaptive estimator that does not depend on R or α and achieves the minimax quadratic risk relative to

$$\mathcal{F}^* = \mathcal{B}_{\alpha,2,\infty}(R)$$

for any fixed, but unknown $R > 0$ and $\alpha \in (0, r)$. As shown by Yang and Barron [1999, page 1591], this minimax risk is of order $n^{-2\alpha/(2\alpha+1)}$. Baraud's estimator may be interpreted as first selecting, based on the data, a "model" $\mathcal{F}_{\hat{k}}$ from the list $\mathcal{F}_0, \mathcal{F}_1, \ldots$, and then estimating the parameters of $\mathcal{F}_{\hat{k}}$ using the least-squares estimator. It always satisfies $\hat{k} \leq J_n$ for some $J_n \leq \log n$, but note that the dimensionality of any model $\mathcal{F}_k$ is exponential in k. For further details, we refer to [Baraud, 2002].

To apply our results to this setting, we need to recast it as a density estimation problem. This is done in the standard manner. For each k, we may parametrise the elements of $\mathcal{F}_k$ by $\theta = (\theta_1, \ldots, \theta_{m_k}) \in \mathbb{R}^{m_k}$, such that $f_\theta \in \mathcal{F}_k$ is

$$f_\theta(X_i) = \sum_{j=1}^{m_k} \theta_j \phi_{(k,j)}(X_i).$$

We then define a corresponding model $\mathcal{M}_k$ as the family of conditional distributions $p_\theta(Y_i \mid X_i)$ that are normal distributions with mean $f_\theta(X_i)$ and variance σ^2. The least squares estimator in the original setting is equivalent to the maximum likelihood estimator in the density estimation setting. Hence the prediction strategy for model $\mathcal{M}_k$ is defined by $p_k(Y_{n+1} \mid x^{n+1}, y^n) := p_{\hat{\theta}_k(x^n, y^n)}(Y_{n+1} \mid x_{n+1})$, where $\hat{\theta}_k(x^n, y^n) \in \mathbb{R}^{m_k}$ is the ML estimator within $\mathcal{M}_k$. (The ML estimator may not be uniquely defined for the first few outcomes, but this can be addressed by using any default prediction strategy p_0 that guarantees a finite risk. Since the ML estimator is uniquely defined almost surely for $n > m_k$, this does not change the asymptotics. We omit the details.) From this perspective Baraud's model selection criterion becomes equal to AIC on a subset $\mathcal{F}_1, \ldots, \mathcal{F}_{J_n}$ of the models, except that it selects $\mathcal{F}_0$ if the norm of the maximum likelihood estimate exceeds a certain threshold.

For regression with independent Gaussian errors, the expected KL divergence is proportional to the expected quadratic (L_2) error:

$$\mathbf{E}_{p_X^*} D(p^*(Y \mid X) \| p_\theta(Y \mid X)) = \frac{1}{2\sigma^2} \mathbf{E}_{p_X^*} (f_\theta(X) - f^*(X))^2.$$

Therefore, the minimax instantaneous quadratic risk is also proportional to the minimax KL risk of the corresponding prediction strategy, and the same holds for the cumulative risks. By Proposition 2.2 this cumulative minimax risk is of order $n \cdot n^{-2\alpha/(2\alpha+1)} = n^{1/(2\alpha+1)}$, and is achieved by Baraud's adaptive least-squares estimator. The corresponding adaptive density estimator thus achieves the same cumulative KL risk, up to a constant factor.

As formulated, Theorem 2.1 cannot be applied to conditional densities, but we can extend it by *deconditioning* as follows: for each $p_\theta(Y|X)$, we define the joint density $p_\theta(X,Y) := p_\theta(Y|X)p_X^*(X)$. The predictions of the switch distribution for the joint densities then also take the form $p_{\mathrm{sw}}(X,Y) = p_{\mathrm{sw}}(Y|X)p_X^*(X)$, where $p_{\mathrm{sw}}(Y|X)$ does not depend on p_X^*. The theorem can then be applied to the models $\mathcal{M}_0', \mathcal{M}_1', \ldots$ of the joint densities. Since, for all joint densities p, $E_{p_X^*} D(p^*(Y|X) \| p(Y|X)) = D(p^*(X,Y) \| p(X,Y))$, Theorem 2.1 also applies to the models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ consisting of the conditional densities, as long as we can equip the switch distribution with a prior for which Condition 2.1 holds.

Since the model $\mathcal{F}_k$ chosen by Baraud's adaptive estimator at sample size n always satisfies $k \leq \log n$ [Baraud, 2002, Eq. 17] we may simply take $\mathcal{L}_n = \{0, \ldots, \log n\}$, $\mathcal{K}_n$ as in (2.17) and the prior π defined as in (2.11). Then Theorem 2.1 implies that the slow switch distribution achieves the same cumulative KL risk $n^{1/(2\alpha+1)}$ as Baraud's adaptive estimator, up to asymptotically the same factor. The minimax cumulative KL risk relative to the Besov balls $\mathcal{B}_{\alpha,2,\infty}(R)$ is also of order $n^{1/(2\alpha+1)}$ [Yang and Barron, 1999, page 1592]. Therefore, the switch distribution achieves the minimax cumulative KL risk in this setting. By defining $\mathcal{K}_n$ as in Theorem 2.2 and using again a uniform prior on $\mathcal{K}_n$, we can also verify that the conditions of Theorem 2.2 hold, and hence that the fast switch distribution also achieves the minimax cumulative KL risk.

We stress that this is just one particular instance of an adaptive estimator to which our theorem can be applied. For example, Baraud's estimator applied to another collection of $\mathcal{F}_k$ based on wavelets leads to the minimax quadratic risk over some Besov balls $\mathcal{B}_{\alpha,l,\infty}(R)$ with $l \geq 1$;

Birgé [2004] extends these results to yet more general settings using robust rather than least-squares estimators; in all these cases, Theorem 2.1 can be used to show that the slow and fast switch distribution, based on the same estimators, achieve the minimax cumulative KL risk.

2.6 Consistency

In Section 2.3.3 we have introduced the model selection criterion δ_{sw}, which selects the model from $\mathcal{L}$ with highest posterior probability under the switch distribution. It is natural to ask whether δ_{sw} is *consistent*, in the sense that it asymptotically selects the true model $\mathcal{M}_{k^*}$ with probability one if the data X^∞ are actually distributed according to a distribution in $\mathcal{M}_{k^*}$.

Ordinary Bayes factor model selection is consistent if the prediction strategies associated with the models are also Bayesian, and if the models are sufficiently distinct in the sense that the corresponding prediction strategies are mutually singular [Barron et al., 1998]. (Two distributions p_1 and p_2 on $\mathcal{X}^\infty$ are mutually singular if there exists a measurable set $A \subseteq \mathcal{X}^\infty$ such that $p_1(A) = 1$ and $p_2(A) = 0$.) To prove consistency of δ_{sw} we require similar conditions, except that the mutual singularity requirement is made somewhat stricter; this is discussed below the theorem.

Theorem 2.3 (Consistency). *Let* $\mathrm{support}(\lambda_1) \subseteq \mathrm{support}(\lambda_2) \subseteq \cdots$ *and assume* $\mathcal{L} = \bigcup_{n=1}^{\infty} \mathrm{support}(\lambda_n)$. *For all* $k \in \mathcal{L}$, *let* p_k *be a Bayesian prediction strategy relative to some parametric model* $\mathcal{M}_k = \{p_\theta \mid \theta \in \Theta_k\}$ *with corresponding prior density* w_k. *Let* p_{sw} *be the switch distribution with prior* π *as in* (2.10). *Suppose the following conditions hold:*

1. *If* $k, k' \in \mathrm{support}(\lambda_{n+1})$, *then* $p_k(X^\infty \mid X^n)$ *and* $p_{k'}(X^\infty \mid X^n)$ *are mutually singular with probability one if* X^n *is distributed according to either* p_k *or* $p_{k'}$.

2. *Let* $B_n^k = \{((t_1, k_1), \ldots, (t_m, k_m)) \in \mathbb{S} \mid t_m \leq n+1, k_m = k\}$ *denote the set of switching parameters that select* p_k *at their last switch, which also occurs no later than* $n+1$. *For all* $k \in \mathcal{L}$, *there should exist an* $n_k \geq 0$ *such that*

$$\sum_{\mathbf{s} \in B_{n_k}^k} \pi(\mathbf{s}) q_{\mathbf{s}}(X^{n_k}) > 0 \qquad (p_k\text{-a.s.}) \tag{2.23}$$

Then, for all $k^* \in \mathcal{L}$, for all $\theta^* \in \Theta_{k^*}$ except for a subset of Θ_{k^*} of w_{k^*}-measure 0, the posterior distribution of the switch distribution on K_{n+1} satisfies

$$p_{\mathrm{sw}}(K_{n+1} = k^* \mid X^n) \stackrel{n\to\infty}{\longrightarrow} 1 \qquad \text{with } p_{\theta^*}\text{-probability 1,} \tag{2.24}$$

which implies consistency of δ_{sw} as defined in (2.14).

For $k \in \mathcal{L}$ such that $\lambda_1(k)$ is positive, (2.23) in the second requirement is trivially satisfied with $n_k = 0$. This is the case case for all $k \in \mathcal{L}$ if the support of λ_n does not depend on n. For $n_k > 0$, the second requirement expresses that if $\lambda_1(k) = 0$, but $\lambda_{n_k+1}(k) > 0$, then there should be some way for the switch distribution to switch to k without giving zero density to the data. This requirement is already satisfied if there is a single prediction strategy p_k with $\lambda_1(k) > 0$ such that $p_k(x_{n+1} \mid x^n) > 0$ for all x^n, x_{n+1}.

Thus the requirements of Theorem 2.3 are primarily about the prediction strategies p_k indexed by $\mathcal{L}$; the second condition is the only constraint on the prediction strategies indexed by $\mathcal{K}$. As such, the consistency theorem applies to the basic version of the switch distribution, as well as to the slow and fast switch distributions of Section 2.5. It is even more widely applicable, as, in contrast to our risk rate results above, it does not require i.i.d. data.

Requirement 1 deserves some further discussion. We first consider ordinary mutual singularity. Consider two Bayesian prediction strategies p_1 and p_2 with priors w_1 and w_2 on parameter spaces Θ_1 and Θ_2 of the corresponding models $\mathcal{M}_1$ and $\mathcal{M}_2$. Then $p_1(X^\infty)$ and $p_2(X^\infty)$ are mutually singular if the models contain stationary ergodic distributions and the induced priors on the space of distributions are mutually singular. This is the case, for example, if the elements of $\mathcal{M}_1$ and $\mathcal{M}_2$ are i.i.d. or Markov distributions, and Θ_1 and Θ_2 are of different dimensionality with priors w_1 and w_2 that are absolutely continuous with respect to Lebesgue measure [Barron et al., 1998, Dawid, 1992b]. Note that this includes the case of nested models $\mathcal{M}_1 \subset \mathcal{M}_2$ that are parametrised in the same way (i.e. $\Theta_1 \subset \Theta_2$), because then the difference in dimension ensures that $w_2(\Theta_1) = 0$.

Thus the requirement that $p_1(X^\infty)$ and $p_2(X^\infty)$ are mutually singular is quite weak. However, we require mutual singularity to hold conditional on almost all initial sequences of outcomes x^n. If $p_1(X^n)$ and $p_2(X^n)$ are equivalent (i.e. either distribution is absolutely continuous with respect to the other), then the posteriors $w_1(\theta \mid X^n)$ and

$w_2(\theta \mid X^n)$ are almost surely well defined and mutual singularity of the priors $w_1(\theta)$ and $w_2(\theta)$ implies mutual singularity of the posteriors, such that Requirement 1 is satisfied under the same weak conditions as were given for mutual singularity of $p_1(X^\infty)$ and $p_2(X^\infty)$. If they are not equivalent, then it matters how $p_1(X_{n+1} \mid x^n)$ and $p_2(X_{n+1} \mid x^n)$ are defined when $p_1(x^n) = 0$ or $p_2(x^n) = 0$. If for all x^n this is done such that $p_1(X^\infty \mid x^n)$ and $p_2(X^\infty \mid x^n)$ are mutually singular, then again Requirement 1 is satisfied under the conditions above.

Thus, the consistency theorem applies in many of the situations where Bayes factor model selection is used [Kass and Raftery, 1995], including, for example, learning of the number of components of a mixture distribution, Markov order estimation (as in the introductory example), histogram density estimation with fixed bin widths (see below) and Gaussian regression with random design. In all these cases, for $k \neq k'$, the models $\mathcal{M}_k$ and $\mathcal{M}_{k'}$ either have empty intersection or are nested but of different dimensionality, which is sufficient for Requirement 1.

2.6.1 Combining Risk Results and Consistency

Although both our cumulative risk theorems and our consistency theorem are quite general, there is one difficulty in applying both at the same time. The risk theorems allow us to piggyback on existing results where an estimator is proved to achieve minimax risk. However, these estimators are often not Bayesian, so that the requirement of Theorem 2.3 is not satisfied. There are three obvious methods to bridge this gap: first, in the current framework the prediction strategies in $\mathcal{K}$ are defined to be frozen versions of the prediction strategies in $\mathcal{L}$; thus $\mathcal{K}$ necessarily contains (frozen versions of) Bayesian estimators if consistency is to be established. However the framework can be relaxed somewhat, by allowing $\mathcal{K}$ to consist of (frozen versions of) non-Bayesian estimators, while $\mathcal{L}$ remains unchanged. A second solution would be to generalise Theorem 2.3 to other estimators (an initial such generalisation is provided by [Van Erven et al., 2008a]).

In the following example we take a third approach: we show that the risk of the Bayesian estimator must be so close to the risk of an estimator that is known to achieve minimax risk, that it must achieve minimax risk itself. In the regression setting of the example above, the Bayesian predictions based on Jeffreys' prior are almost identical

to ML predictions (Section 2.5.5). In similar fashion it is possible to establish minimax cumulative risk as well as consistency for histogram density estimation, where the models $\mathcal{M}_k$ are regular, fixed-bin width histograms (as in, e.g., [Rissanen et al., 1992]).

Example 2.5. We now show that both our results on achieving the minimax cumulative risk as well as our consistency theorem can be applied to Gaussian linear regression with random i.i.d. design. We consider Gaussian models $\mathcal{M}_0, \mathcal{M}_1, \ldots$ based on linear combinations of orthonormal bases $S_0, S_1, \ldots$ as described in Section 2.5.5, where each model $\mathcal{M}_k$ is represented by a Bayesian estimator p_k^{JP} based on Jeffreys' prior. The result holds for general bases $S_0, S_1, \ldots$, not just those considered by Baraud [2002].

Let $\phi_k(x) = (\phi_{(k,1)}(x), \ldots, \phi_{(k,m_k)}(x))$ and let $\Phi_k = (\phi_k(x_1)^T, \ldots, \phi_k(x_n)^T)^T$ be the $n \times m_k$ design matrix (see, e.g. [Grünwald, 2007, page 357]). The Bayesian prediction strategies p_k^{JP} are similar to the prediction strategies $p_k(Y_{n+1} \mid x^{n+1}, y^n)$ based on the ML estimator $\hat{\theta}_k(x^n, y^n)$, as defined in Section 2.5.5. In both cases the predictive distribution is Gaussian with the same mean $\phi_k(x_{n+1})^T \hat{\theta}_k(x^n, y^n)$. But whereas the variance for the ML estimator p_k is σ^2, the variance of p_k^{JP} is $\sigma^2(1 + \phi_k(x_{n+1})^T(\Phi_k^T \Phi_k)^{-1}\phi(x_{n+1}))$. As with the ML estimator, p_k^{JP} is not uniquely defined for the first few outcomes, for which we have to substitute a default strategy p_0 that guarantees a finite risk. If the design matrix is almost surely invertible, as is implied by the assumption in Baraud's result that p_X^* has a density relative to Lebesgue measure, this does not change the asymptotics.

Theorem 2.3 extends to conditional densities by deconditioning as in Section 2.5.5. Its requirements are now satisfied: mutual singularity follows from the i.i.d. setup, and because all predictive densities are positive, Requirement 2 is also trivially satisfied. We thus obtain consistency of the switch distribution for regression with p_k^{JP}.

At the same time, switching based on p_k^{JP} achieves the minimax risk in the setting of Baraud that we described in Section 2.5.5. In that section, we already indicated that switching with ML-based p_k achieves the minimax risk. Minimaxity of switching based on p_k^{JP} follows because of the following fact: let $(X_1, Y_1), (X_2, Y_2), \ldots$ be as specified in the beginning of Section 2.5.5. Then for any k and n such that $\hat{\theta}_k(X^n, Y^n)$ exists

almost surely, the KL risk of p_k^{JP} is no larger than the risk of the p_k:

$$r(p^*, p_k^{\mathrm{JP}}, n+1) \leq r(p^*, p_k, n+1).$$

This follows from calculations based on the relations between KL divergence and squared error as in [Grünwald, 2007, Chapter 12]. We omit the details.

2.7 Simulation Study

In order to test the switch distribution as a general tool for model selection and prediction, we consider sequential polynomial regression on simulated data. The general setup is as in Section 2.5.5; but instead of Baraud's instantiation we use $\mathcal{A} = [-1, 1]$ and $S_k = \{x^0, x^1, \ldots, x^k\}$, $\mathcal{F}_k$ being the corresponding space of linear combinations of S_k, i.e. the set of all k-degree polynomials, and $\mathcal{M}_k$ being the corresponding conditional densities. We take a fixed variance $\sigma^2 = 1$. As in Example 2.5, we associate Bayesian prediction strategies $p_0, p_1, \ldots$ with the models, with Jeffreys' prior on the model parameters.

We consider polynomials of order 0 up to a fixed maximum order K. Six methods are evaluated: $\mathcal{C} = \{$Fast switch, Slow switch, Basic switch, Bayes, AIC, BIC$\}$. With each method, we associate a model selection criterion and an estimator.

The switch distribution is defined as in Section 2.2.2; the associated model selection criterion is given by (2.14). We used $\mathcal{L}_n = \mathcal{L} = \{0, 1, \ldots, K\}$ and three different definitions of $\mathcal{K}_n$: for the basic switch distribution we have $\mathcal{K}_n = \mathcal{L}$; for the slow and fast switch distributions, $\mathcal{K}_n$ is defined as in (2.17) and Theorem 2.2, respectively. In case of the fast switch distribution, the prediction strategies were frozen at each distinct value of $\left\lfloor 1.1^i \right\rfloor$ for $i = 0, 1, 2, \ldots$ The priors are chosen as in (2.11), where the supports of λ_n and κ_n are $\mathcal{L}_n$ and $\mathcal{K}_n$, respectively.

The Bayesian method uses a uniform prior on the models; the model that maximises the a posteriori probability is selected. Prediction proceeds using model averaging, where the models are weighted according to their posterior probabilities.

The AIC and BIC criteria associate values v_k with the order k polynomial models; for AIC this is $v_k = -\ln \hat{p}_k + (k+1)$ and for BIC $v_k = -\ln \hat{p}_k + \frac{1}{2}(k+1)\ln n$, where $\hat{p}_k = \max\{p(y^n \mid x^n) \mid p \in \mathcal{M}_k\}$

is the maximum likelihood of the data using the order k polynomial model. The model k selected by AIC or BIC is the one that minimises v_k; while for AIC and BIC prediction is often done using the selected model only, to obtain competitive results it is necessary to use a mixture of $p_0, \ldots, p_K$, as proposed by Akaike [1979]. Thus, for AIC and BIC the predictions $\{p_k(Y_{n+1}|x^{n+1}, y^n) \mid k \in \mathcal{L}_n\}$ are weighted using $w_k = \exp(-v_k) / \sum_{k=0}^{K} \exp(-v_k)$.

We have subjected these model selection criteria to a simulation experiment which is most easily expressed in the form of an algorithm. As input it takes a "true" regression function $f^* : [-1, 1] \to \mathbb{R}$, the number of outcomes N to be predicted, the maximal model order K and the number of runs R.

Algorithm 2.2 TEST(f^*, N, K, R)

for $r = 1, \ldots, R$ **do**
  for $n = 1, \ldots, N$ **do**
    for $c \in \mathcal{C}$ **do**
      Ask criterion c to select a model $k \in \mathcal{L}$
      Sample x_n uniformly at random from $[-1, 1]$.
      Ask criterion c to form prediction $p(Y_n \mid x^n, y^{n-1})$.
      Sample y_n from a normal density with mean $f^*(x_n)$ and variance 1.
      Accumulate individual sequence redundancy $\log_2 \left(\frac{\varphi(y_n - f^*(x_n))}{p\left(y_n \mid x^n, y^{n-1}\right)} \right)$, where φ is the standard normal density
    end for
  end for
end for

By subsequently averaging the results from the R runs, we obtain estimates of the mean selected model and of the cumulative risk as a function of the number of observations for each method.

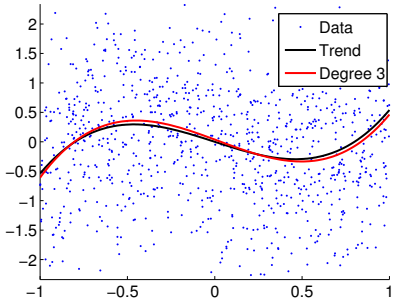
We ran the testing algorithm with the following two sets of parameters:

1. $f^*(x) = 1.5x^3 - 0.96x$; $R = 200$, $N = 1000$ and $K = 6$.
2. $f^*(x) = 2$ if $x \in [-\frac{1}{2}, \frac{1}{2}]$ and -2 otherwise; $R = 50$, $N = 600$ and $K = 35$.

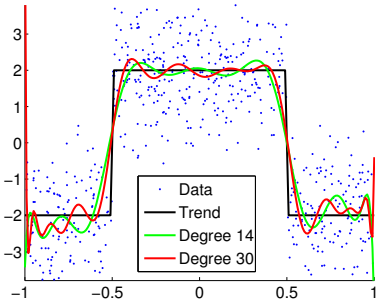
(For the slow switch distribution we used a reduced value of N, in order to obtain a running time comparable to that of the other criteria.) In the first experiment, the generating distribution is in $\mathcal{M}_3$ (the set of third degree polynomials with standard normal noise), so we are in a parametric scenario where consistency is relevant. In the second experiment, the true distribution is not in any of the models, but it can be arbitrarily well approximated by polynomials, a prototypical nonparametric scenario.

Results The left column of Figure 2.3 shows the results for the first experiment, the right column for the second experiment. The first row shows an example data set, together with f^* and an example fit for one or two reasonable models. The second row shows the average index of the selected model for each criterion. The third row shows the estimated cumulative risk (measured in bits), with an indication of the standard error of the estimate (standard deviation of the individual runs divided by $\sqrt{R}$).

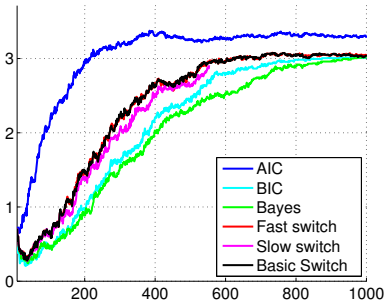
In the parametric case, we would expect Bayes, BIC and all versions of the switch distribution to consistently select a degree of 3 for sufficiently large sample sizes. This is confirmed by the results, but note in Figure 2.3c that Bayes and BIC appear to require a larger sample on average before detecting that $\mathcal{M}_3$ is true. Also, the slow switch distribution seems to select models of a slightly lower order than the two fast varieties of switching. Finally, the AIC criterion is by far the most responsive: it is substantially quicker to determine that at least a degree 3 polynomial is required to obtain the best predictions; on the other hand even after a lot of data have become available, AIC often selects a polynomial order larger than three, as it is inconsistent. In Figure 2.3e we see that generally, the quicker a method is to detect when the third degree polynomial model starts making the best predictions, the smaller its cumulative risk. Thus, AIC is a clear winner, followed by the fast and basic switch distributions, then the slow switch distribution, and finally BIC and Bayes. The more conservative behaviour of the latter two methods is explained by the occurrence of the catch-up phenomenon. Interestingly, over roughly the first 100 outcomes AIC actually performs *worst*: it starts selecting higher order models even before the instantaneous risk for those models drops below that for lower order models. Possibly this effect can be mitigated using a small sample



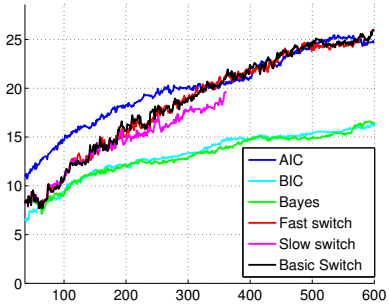
(a) Typical data



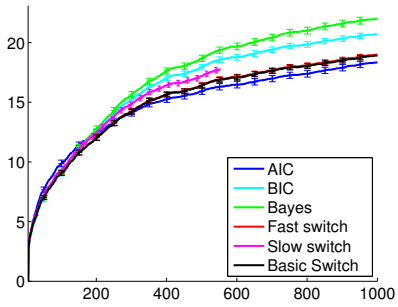
(b) Typical data



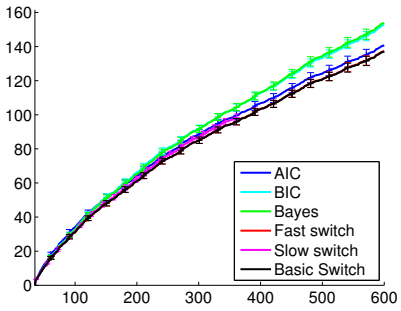
(c) Average degree selected



(d) Average degree selected



(e) Estimated cumulative risk



(f) Estimated cumulative risk

Figure 2.3: Sequential polynomial regression results

correction for AIC, such as in AIC_c [Burnham and Anderson, 2002].

In this parametric experiment, eventually all consistent methods select $\mathcal{M}_3$, so their instantaneous risks converge to the instantaneous risk of p_3. Thus, the difference in cumulative risk for these methods will converge to a constant. In fact, by $n = 1000$ the lines for each method already appear to run more or less parallel. Empirically, AIC seems to follow the same trend; it is unclear whether its cumulative risk has the same asymptotics.

In the nonparametric case (the right column of Figure 2.3), we observe an even greater discrepancy in the model order selected by BIC and Bayes compared to the methods that do not suffer from the catch-up phenomenon. As in the parametric case, AIC initially selects models of an overly high order, for which it is punished slightly in terms of cumulative risk. From $n = 300$ onwards AIC and the switch distributions seem to be in approximate agreement on the best model order, whereas Bayes and BIC lag behind dramatically. As a result, the differences in cumulative risk for these methods are substantially larger than in the parametric experiment.

Interpretation The experiments confirm the theoretical results of the chapter: (1) all considered methods except AIC are consistent, (2) BIC and Bayes suffer from the catch-up phenomenon and as such issue inferior predictions. The predictive performance of the switch distribution, at least in its fast and basic incarnations, is competitive with AIC.

Note that the cumulative risk for all methods is actually quite small in these particular experiments: only about 20 bits in the parametric case. Because of this, the size of $\mathcal{K}_n$, which determines the overhead of switching, can have a substantial effect on the results. This is probably why the slow switch distribution appears to be more "sluggish" in switching to higher order models than the fast and basic switch distributions: since $\mathcal{K}_n$ contains substantially more prediction strategies for the slow switch distribution than for the other two variants, the prior probability $\kappa_n(k) = 1/|\mathcal{K}_n|$ of switching to a particular estimator p_k will be correspondingly lower.

This is clearly an issue that deserves careful consideration in practice if the cumulative risk is very small. Whether or not it is small depends very much on the setting; recall that in the Markov chain example in the introduction a single switch yielded a reduction in code

length of about 7 000 bits. Compared to this the overhead induced by a couple of switches is negligible. Even when the cumulative risk is very small, it still cannot do much harm to use the switch distribution; for the prior used in these experiments the cumulative risk of the switch distribution is at most one bit more than that of Bayes (see Example 2.3).

2.8 Discussion

In this section we put our results in a broader perspective. First we discuss the AIC-BIC dilemma in more detail. Then we consider two alternative criteria of predictive performance that one might be interested in: first, how well does the switch distribution predict when only the model with highest posterior probability is used for prediction, instead of a mixture? Second, our analysis is in terms of the minimax cumulative risk; to what extent do our results carry over to the instantaneous risk setting? Then, since most of our results about cumulative risk are for the nonparametric setting, we compare our approach to the nonparametric Bayesian methods that have proved to be quite effective in recent years. Finally, we indicate a number of areas where our results might be strengthened in future research.

2.8.1 The AIC-BIC Dilemma

Over the last 25 years or so, the question of whether to base model selection on AIC or BIC type methods has received a lot of attention in the theoretical and applied statistics literature, as well as in fields such as psychology and biology, where model selection plays an important role [Speed and Yu, 1993, Hansen and Yu, 2001, 2002, Barron et al., 1994, Forster, 2001, de Luna and Skouras, 2003, Sober, 2004]. It has even been suggested that, since these two types of methods have been designed with different goals in mind (optimal prediction vs "truth hunting"), it may simply be the case that *no* procedures exist that combine the best of both types of approaches [Sober, 2004]. Still, for practitioners, the incompatibility of the two methods remains worrying. Consider, for example, a psychologist who wants to determine how some response Y (e.g., reaction times in a memory experiment) depends on input variables X and Z (e.g. gender and age). He models Y as a sum of a linear function of X and a polynomial of Z. Now according to some statisti-

cians, we are supposed to tell the psychologist: if you use an AIC-type method, you need fewer data to learn a model that predicts well. But, in case Y is independent of X, then you may not find out, even if you do have a lot of data. On the other hand, if you use a BIC-type method, the situation is reversed. Thus, you should first determine what your goal is — finding out about independency or prediction — and only then can I tell you what method to use. The problem with this is that in practice, the psychologist's main goal is often neither predictive optimality nor consistency; so he cannot tell. He just wants a method that gives useful insight into the structures underlying the data, and he wants to use this insight to guide his further research. To gain confidence that the chosen method will do a good job towards this inherently vague goal, he would like the method to satisfy as many sanity checks as possible. Thus, consistency and predictive optimality play the role of sanity checks rather than direct goals, and we feel that *if* a method exists that satisfies both checks, then this may be a good method for the practitioner to use.

Now, if the AIC-BIC dilemma is interpreted as a conflict between consistency and optimal sequential prediction, then cumulative risk is a natural and often considered performance criterion Haussler and Opper [1997], Rissanen et al. [1992], Barron [1998], Yang and Barron [1999] and Poland and Hutter [2005], and we can reasonably claim that our results solve the dilemma. However it can also be interpreted as a dichotomy between model selection for truth finding and model selection-based (nonsequential) estimation. In that case we do leave a number of loose ends that are discussed in Sections 2.8.2 and 2.8.3.

2.8.1.1 Earlier Approaches

Several other authors have provided procedures which have been designed to behave like AIC whenever AIC is better, and like BIC whenever BIC is better; and which empirically seem to do so. These include *model meta-selection* [de Luna and Skouras, 2003, Clarke, 1997], and Hansen and Yu's *gMDL* version of MDL regression [Hansen and Yu, 2001]; also the "mongrel" procedure of Wong and Clarke [2004] has been designed to improve on Bayesian model averaging for small samples. Compared to these other methods, ours seems to be the first that *provably* is both consistent and minimax optimal in terms of cumulative risk, for some classes $\mathcal{M}^*$. The only other procedure that we

know of for which somewhat related results have been shown, is a version of cross-validation proposed by Yang [2007a] to select between AIC and BIC in regression problems. Yang shows that a particular form of cross-validation will asymptotically select AIC in case the use of AIC leads to better predictions, and BIC in the case that BIC leads to better predictions. In contrast to Yang, we use a single paradigm rather than a mix of several ones (such as AIC, BIC and cross-validation) — essentially our paradigm is just that of universal individual-sequence prediction, or equivalently, the individual-sequence version of predictive MDL, or again equivalently, Dawid's prequential analysis applied to the log scoring rule. Indeed, our work has been heavily inspired by prequential ideas. In [Dawid, 1992a] it is already suggested, without giving any details, that model selection should be based on the *transient* behaviours in terms of sequential prediction of the estimators for the models: one should select the model that is optimal at the given sample size, and this will change as more data become available.

2.8.2 Model Selection vs Model Averaging

In model selection, we are usually given a batch sample at some fixed sample size n and have to choose one (or a few) models. For example, a scientist such as the psychologist above may ask a statistician to advise on a good model. Suppose the statistician advises to use a particular model. The scientist and his colleagues may then adopt this model as a working hypothesis, and use it to make predictions about future data, using some estimator defined relative to the chosen model. As it is unrealistic to switch between models with each new observation, they will tend to use the same model for a while.

If selecting a low-risk model is the goal, then two issues crop up. First, our risk convergence results only apply when predictions are allowed to be a mixture of the predictions of the models, but this may be impractical. One may therefore prefer a model selection criterion that uses the predictions of a single model only. It is quite possible that an analogue of the results in Section 2.5 still holds in this situation; establishing whether it does is posed as an open problem in Section 2.8.5.

Second, in the model selection setting, the instantaneous risk at sample size n, rather than the cumulative risk, is the relevant quantity, since it will determine the quality of the scientist's predictions for several future samples $x_{n+1}, x_{n+2}, \ldots$ In the following subsection we discuss to

what extent our results transfer to instantaneous risk.

2.8.3 Cumulative vs Instantaneous Risk

In the parametric case, based on Theorem 2.3 and the discussion in Section 2.4.4, the switch distribution is consistent under mild conditions, and achieves the minimax cumulative risk. However, an intriguing result was obtained by Yang [2005], who shows that there are scenarios in linear regression where no model selection or model combination criterion can be both consistent and achieve the minimax rate of convergence; Yang [2007b, Theorem 3] gives an explicit lower bound on the factor by which consistent model selection procedures must miss the minimax rate in a simple linear regression problem. In other words, there are parametric scenarios where it is possible, quite straightforward even, to achieve minimax cumulative risk while retaining consistency, whereas minimax instantaneous risk is impossible to achieve without losing consistency. In such cases, clearly, the switch distribution does not achieve minimax instantaneous risk.

Let us nevertheless compare instantaneous risk to cumulative risk for fixed p^*. As shown in [Grünwald, 2007], instantaneous risk convergence is a stronger notion than cumulative risk convergence: for example, suppose we are in the nonparametric setting and the instantaneous risk satisfies $r(p^*, p, n) \preceq cn^{-\gamma}$, then one can easily verify that the average cumulative risk satisfies $n^{-1}R(p^*, p, n) \preceq cn^{-\gamma}$. The converse does not hold: clearly, the instantaneous risk may be larger than the average cumulative risk for some n. However [Grünwald, 2007, Theorem 15.2, page 473], the *gap* between any two n and $n' > n$ at which the risk of p exceeds $cn^{-\gamma}$ must grow without bound as n increases. Thus, small cumulative risk implies small instantaneous risk at "most" sample sizes.

Perhaps more significantly, in the nonparametric case a simple modification of the switch distribution actually achieves minimax *instantaneous* risk, whenever the switch distribution itself achieves the minimax *cumulative* risk. Let p_{sw} be the fast or the slow switch distribution of Sections 2.5.3 and 2.5.4, and define the time average of the switch distribution as

$$\bar{p}_{\mathrm{sw}}(X_n = x, K_n = k \mid x^{n-1}) := \frac{1}{n}\sum_{i=1}^{n} p_{\mathrm{sw}}(X_i = x, K_i = k \mid x^{i-1}),$$

so that the corresponding predictive distribution satisfies

$$\begin{aligned}\bar{p}_{\mathrm{sw}}(X_n = x \mid x^{n-1}) &= \sum_{k \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} p_{\mathrm{sw}}(X_i = x, K_i = k \mid x^{i-1}) \\ &= \frac{1}{n} \sum_{i=1}^{n} p_{\mathrm{sw}}(X_i = x \mid x^{i-1}).\end{aligned}$$

We have the following result, proved in Section 2.9.3:

Proposition 2.3. *Suppose $\mathcal{M}^*$ is a standard IID model class that satisfies Condition 2.2 with constants c_1 and c_2, and $r_{\mathrm{mm}}(n) < \infty$ for all n. If $R_{\mathrm{m}}(p_{\mathrm{sw}}, n) \preceq c_3 \, R_{\mathrm{mm}}(n)$ for a constant c_3, then*

$$r_{\mathrm{m}}(\bar{p}_{\mathrm{sw}}, n) \preceq \frac{c_2}{c_1} \frac{c_3}{1-\gamma} r_{\mathrm{mm}}(n).$$

Note that, if $x_1, x_2, \ldots$ are such that for some fixed k^*, $p_{\mathrm{sw}}(K_n = k^* \mid x^n) \to 1$ as $n \to \infty$, then by definition of $\bar{p}_{\mathrm{sw}}$, we must also have that $\bar{p}_{\mathrm{sw}}(K_n = k^* \mid x^n) \to 1$. Hence, consistency of the switch distribution implies consistency of the time-averaged switch distribution. Consequently, under the appropriate conditions, the time-averaged switch distribution resolves the following version of the AIC-BIC: it is consistent in the parametric case, and achieves the minimax instantaneous risk in the nonparametric case. Since, intuitively, $\bar{p}_{\mathrm{sw}}$ learns (much) "more slowly" than p_{sw}, we suspect that when Condition 2.2 applies, p_{sw} also achieves the minimax instantaneous risk, and hence also resolves this version of the AIC-BIC dilemma.

2.8.4 Nonparametric Bayes

Our results mostly apply to nonparametric inference, where the true distribution is not assumed to be a member of a parametric model. In practice, Bayesian model averaging on a set of parametric models is often used in such scenarios, but a subjective Bayesian should not be surprised that this gives suboptimal results, since under the standard hierarchical prior used in p_{bma} (first a discrete prior on the model index, then a density on the model parameters), we have that with prior-probability 1, p^* is "parametric", i.e. $p^* \in \mathcal{M}_k$ for some k. Thus from the subjective perspective, the hierarchical prior is not really suitable for the situation that we are trying to model, and one should use

a nonparametric prior instead. Indeed, nonparametric Bayesian methods have become very popular in recent years, and they often work very well in practice. Still, their practical and theoretical performance strongly depend on the used priors, and it is often far from clear what prior to use in what situation. In some situations, certain nonparametric priors achieve optimal rates of convergence, but others can even make Bayes inconsistent [Diaconis and Freedman, 1986, Grünwald, 2007].

In minimum description length inference, there are no philosophical objections to doing nonparametric inference using parametric models. In fact, approximating nonparametric families by sequences of finite dimensional parametric models is a standard approach [Barron and Cover, 1991]. Consequently, we view the switch distribution as an MDL method, even though its definition is compatible with the Bayesian framework. Apart from choosing a reasonable sequence of parametric models, it does not require any difficult modelling decisions. Nevertheless, under reasonable conditions the switch distribution achieves the minimax cumulative risk in nonparametric settings, while at the same time, in the words of Barron and Cover, "we retain the possibility of delight in the discovery of the correct family in the finite-dimensional case".

2.8.5 Future Work

We conclude the discussion by suggesting three directions in which our results might be extended.

Other Ways to deal with Increasing Risk - non-i.i.d. settings The "fast" and "slow" versions of the switch distribution differ in their selection of frozen strategies in the definition of $\mathcal{K}_n$. The basic switch distribution uses $\mathcal{K}_n = \mathcal{L}_n$, which works well in practice but invalidates the proofs of Theorems 2.1 and 2.2. It seems unlikely to us that increasing risk would harm performance of the switch distribution too much in practice. The question thus becomes: is there a reasonable assumption one can make about how much the risk is allowed to grow, so that an analogue of Theorem 2.1 can be shown for the basic switch distribution with $\mathcal{K}_n = \mathcal{L}_n$? Relatedly, the basic switch distribution was shown in the introduction to empirically behave very well in a non-i.i.d. setting, a setting that our current risk convergence theorems cannot deal

with. Dealing with increasing risk may also allow one to extend the convergence rate theorems to non-i.i.d. settings.

Predictive Performance in the Model Selection Setting It is unclear whether there is an analogue of our cumulative risk theorems for model *selection* rather than averaging. For example, in Figure 2.1, sequentially predicting using the prediction strategy $p_{\delta_{\mathrm{sw}}(x^n)}$ for the model with index $\delta_{\mathrm{sw}}(x^n)$, which has maximum a posteriori probability (MAP) under the switch distribution, is only a few bits worse than predicting by model averaging based on the switch distribution, and still outperforms standard Bayesian model averaging by about 7 200 bits. However, it is unclear whether or not prediction based on selecting a single model will always perform this well. Analogous results in the MDL literature suggest that a theorem bounding the risk of switch-based model selection, if it can be proved at all, would bound the squared Hellinger rather than the KL risk [Grünwald, 2007, Chapter 15].

Exponentially Many Models Because of Condition 2.1, our theoretical results do not cover the case in which $|\mathcal{L}_n|$, the number of considered models, is exponential in the sample size. Yet this case is very important in practice, for example in the variable selection problem [Shibata, 1983, Li, 1987, Yang, 1999], where at sample size n one considers all 2^n possible subsets of n variables. In such cases AIC is known to lead to severe overfitting [Yang, 1999], and is therefore not suitable.

As it seems clear that the catch-up phenomenon will also occur in model selection problems with exponentially many models, it is an interesting open question whether, for suitable priors λ and κ, the switch distribution can achieve the minimax cumulative risk. To make the method practical, one would then also have to address the computational issues that arise with so many models. Finally, the relation with the popular and computationally efficient L_1-approaches to model selection [Tibshirani, 1996] is as yet also unclear.

2.9 Cumulative Risk Proofs

This section gives the proofs of Theorems 2.1 and 2.2 in Section 2.5, and of Propositions 2.2 and 2.3.

2.9.1 Oracle Approximation Lemma

The proofs of Theorems 2.1 and 2.2 both depend on the following bound on the excess cumulative risk of the switch distribution compared to any oracle.

Lemma 2.1 (Oracle Approximation Lemma)**.** *Let p_{sw} be the switch distribution defined with respect to a prior π (that can be written in the form (2.10)). Let $\mathcal{M}^*$ be a standard IID model class, and let ω be an oracle relative to $\mathcal{K}_1, \mathcal{K}_2, \ldots$ Finally, let $m(n)$ be the maximum number of different prediction strategies that ω uses before the n-th outcome, i.e.*

$$m(n) = \max_{p^* \in \mathcal{M}^*} \max_{x^n \in \mathcal{X}^n} \left| \{ i : 2 \leq i \leq n, \omega(p^*, x^i) \neq \omega(p^*, x^{i-1}) \} \right| + 1. \tag{2.25}$$

We then have, for any $p^* \in \mathcal{M}^*$,

$$R(p^*, p_{\mathrm{sw}}, n) - R(p^*, p_\omega, n) \leq L_{\mathrm{m}}(m(n)+1) + m(n)\Big(L_{\mathrm{k}}(n) + L_{\mathrm{t}}(n+1)\Big),$$

where

$$\begin{aligned} L_{\mathrm{m}}(m) &= \max\{ -\log \mu(a) \mid 1 \leq a \leq m \quad \} \\ L_{\mathrm{t}}(n) &= \max\{ -\log \tau(t) \mid 1 < t \leq n \quad \} \\ L_{\mathrm{k}}(n) &= \max\{ -\log \kappa_t(k) \mid k \in \mathcal{K}_t, 1 \leq t \leq n \}. \end{aligned}$$

Since this holds uniformly for all $p^* \in \mathcal{M}^*$, we also have

$$R_{\mathrm{m}}(p_{\mathrm{sw}}, n) - R_{\mathrm{m}}(p_\omega, n) \leq L_{\mathrm{m}}(m(n)+1) + m(n)\Big(L_{\mathrm{k}}(n) + L_{\mathrm{t}}(n+1)\Big).$$

The bound of the lemma may be interpreted as a uniform bound on the number of bits required to encode how ω switches between prediction strategies. Note that in particular, if π satisfies Condition 2.1, then

$$L_{\mathrm{m}}(m(n)+1) + m(n)\Big(L_{\mathrm{k}}(n) + L_{\mathrm{t}}(n+1)\Big) = O\Big(m(n) \log n\Big).$$

Proof. For arbitrary $p^* \in \mathcal{M}^*$ and $x^n \in \mathcal{X}^*$, let m denote the number of different prediction strategies $k'_1, \ldots, k'_m$ selected by the oracle ω to predict x^n, and let $1 = t'_1 < t'_2 < \cdots < t'_m$ denote the sample sizes at which ω switches between them. That is,

$$t'_j = \min\Big\{ i \mid t'_{j-1} < i \leq n, \omega(p^*, x^i) \neq \omega(p^*, x^{i-1}) \Big\}$$

for $j = 2, \ldots, m$, and $k'_j = \omega(p^*, t'_j)$ for $j = 1, \ldots, m$.

Because ω selects its predictions from $\mathcal{K}_1, \mathcal{K}_2, \ldots$, the switch distribution puts positive prior probability on switch sequences $\mathbf{s}$ such that $q_{\mathbf{s}}(x^n) = p_\omega(x^n)$, where $q_{\mathbf{s}}$ is as in (2.6). Let

$$\mathcal{S} = \{((t_1, k_1), \ldots (t_{m+1}, k_{m+1})) \in \mathbb{S} \mid (t_j, k_j) = (t'_j, k'_j) \text{ for } 1 \leq j \leq m, t_{m+1} = n+1\}$$

denote a convenient subset of these sequences, in which the last switch (at switch-point t_{m+1}) occurs immediately after the n-th outcome. As

$$p_{\mathrm{sw}}(x^n) = \sum_{\mathbf{s} \in \mathbb{S}} q_{\mathbf{s}}(x^n)\pi(\mathbf{s}) \geq \sum_{\mathbf{s} \in \mathcal{S}} q_{\mathbf{s}}(x^n)\pi(\mathbf{s}) = p_\omega(x^n)\pi(\mathcal{S}),$$

our plan is to find a uniform lower bound c on $\pi(\mathcal{S})$, which does not depend on p^* or x^n, and then apply Proposition 2.1 to obtain the desired result. Using that π is of the form (2.10), we see that

$$\begin{aligned}
\pi(\mathcal{S}) &= \sum_{k_{m+1}} \mu(m+1) \left(\prod_{j=1}^{m} \kappa_{t_j}(k_j) \tau(Z = t_{j+1} \mid Z > t_j) \right) \lambda_{t_{m+1}}(k_{m+1}) \\
&= \mu(m+1) \left(\prod_{j=1}^{m} \kappa_{t_j}(k_j) \tau(Z = t_{j+1} \mid Z > t_j) \right) \\
&\geq \mu(m+1) \left(\prod_{j=1}^{m} \kappa_{t_j}(k_j) \tau(Z = t_{j+1}) \right).
\end{aligned}$$

Hence

$$-\log \pi(\mathcal{S}) \leq L_{\mathrm{m}}(m(n)+1) + m(n)\Big(L_{\mathrm{k}}(n) + L_{\mathrm{t}}(n+1)\Big) =: -\log c,$$

and the lemma follows by Proposition 2.1. □

2.9.2 Proof of Theorem 2.1

Proof. Let $1 = t_1 < t_2 < \ldots$ be a sequence of switch-points. We will construct an oracle ω' (relative to $\mathcal{K}_1, \mathcal{K}_2, \ldots$) that switches only at $t_2, t_3, \ldots$ and is such that

$$R_{\mathrm{m}}(p_{\omega'}, n) \preceq R_{\mathrm{m}}(p_\omega, n) \cdot \limsup_{j \to \infty} \frac{d_j}{d_{j-1}}, \tag{2.26}$$

where $d_j = t_{j+1} - t_j$. This construction will work for any choice of switch-points. In particular, by choosing the switch-points such that $d_j = \left\lceil \exp\left(j^{1/(1+\alpha)}\right) \right\rceil$, we obtain

$$\begin{aligned}\limsup_{j\to\infty} \frac{d_j}{d_{j-1}} &= \limsup_{j\to\infty} \exp\left(\frac{j}{j^{\alpha/(1+\alpha)}} - \frac{j-1}{(j-1)^{\alpha/(1+\alpha)}}\right) \\ &\leq \limsup_{j\to\infty} \exp\left(\frac{1}{j^{\alpha/(1+\alpha)}}\right) = 1.\end{aligned}$$

Let $m(n)$ denote the maximum number of different prediction strategies used by ω' before time n, as defined in (2.25). We must have $t_{m(n)} > n$. Hence $m(n) \leq k$ for the smallest k such that $d_k = \left\lceil \exp\left(k^{1/(1+\alpha)}\right) \right\rceil > n$. Solving for k, we obtain $m(n) \leq (\log n)^{1+\alpha}$, which by the Oracle Approximation Lemma implies that

$$R_{\mathrm{m}}(p_{\mathrm{sw}}, n) = R_{\mathrm{m}}(p_{\omega'}, n) + O((\log n)^{2+\alpha}).$$

Together with (2.26) and the assumption that $(\log n)^{2+\alpha} / R_{\mathrm{m}}(p_\omega, n) \to 0$, the conclusion of the theorem follows.

It remains to exhibit the oracle ω' that satisfies (2.26). To this end we first construct an intermediate oracle ω'' (relative to $\mathcal{K}_1, \mathcal{K}_2, \ldots$) whose risk is nonincreasing and never exceeds the risk of ω. Let $s(p^*, n) = \arg\min_{1\leq s\leq n} r(p^*, p_\omega, s)$ denote the sample size at which ω achieved minimal risk before sample size n (ties may be broken arbitrarily). Then for any p^*, n and data x^{n-1}, ω'' is defined as

$$\omega''(p^*, x^{n-1}) = \omega(p^*, x^{s(p^*,n)-1}) \circ s(p^*, n),$$

where $x^{s(p^*,n)-1}$ is the prefix of x^{n-1} of length $s(p^*, n) - 1$. Thus, at sample size n, ω'' copies the prediction made by ω at sample size $s(p^*, n)$, which is possible because that prediction strategy is still available as a frozen strategy. Because p^* is i.i.d. by assumption, the construction guarantees that $r(p^*, p_{\omega''}, n) = r(p^*, p_\omega, s(p^*, n))$, such that the risk of ω'' is nonincreasing and never exceeds the risk of ω.

We proceed to construct the oracle ω' satisfying (2.26). It is defined by copying the predictions of ω'' at the last switch-point. That is, if i is such that $t_j \leq i < t_{j+1}$, then $\omega'(p^*, x^{i-1}) = \omega''(p^*, x^{t_j-1})$. As the

predictions of ω' do not change between switch-points, its risk does not change either, and $r(p^*, p_{\omega'}, i) = r(p^*, p_{\omega''}, t_j)$ for any $p^* \in \mathcal{M}^*$.

Let $c = \limsup_{j\to\infty} d_j/d_{j-1}$ and let $\varepsilon > 0$ be arbitrary. Then there exists a j^* such that $\sup_{j\geq j^*} d_j/d_{j-1} \leq c + \varepsilon$. Now for any n, let m_n be such that $t_{m_n} \leq n < t_{m_n+1}$. Because the risk of ω'' is nonincreasing, we can underestimate its cumulative risk by

$$\sum_{i=1}^{t_{(j^*-1)}} r(p^*, p_{\omega''}, i) \geq \sum_{j=1}^{j^*-1} d_{j-1} r_j,$$

$$\sum_{i=t_{(j^*-1)}+1}^{n} r(p^*, p_{\omega''}, i) \geq \sum_{i=t_{(j^*-1)}+1}^{t_{m_n}} r(p^*, p_{\omega''}, i) \geq \sum_{j=j^*}^{m_n} d_{j-1} r_j,$$

where $r_j = r(p^*, p_{\omega''}, t_j)$ and we define $d_0 = 1$. We can overestimate the cumulative risk of the derived oracle ω' by a similar bound:

$$\sum_{i=1}^{t_{j^*}-1} r(p^*, p_{\omega'}, i) = \sum_{j=1}^{j^*-1} d_j r_j,$$

$$\sum_{i=t_{j^*}}^{n} r(p^*, p_{\omega'}, i) \leq \sum_{i=t_{j^*}}^{t_{(m_n+1)}-1} r(p^*, p_{\omega'}, i) = \sum_{j=j^*}^{m_n} d_j r_j.$$

If $R_{\mathrm{m}}(p_\omega, n) = \infty$ from some n onwards, then the theorem is trivially true, so assume without loss of generality that $R_{\mathrm{m}}(p_\omega, n) < \infty$ for all n, which implies that $\sup_{p^*} \sum_{i=1}^{t_{(j^*-1)}} r(p^*, p_{\omega''}, i) = R_{\mathrm{m}}(p_{\omega''}, t_{(j^*-1)}) \leq R_{\mathrm{m}}(p_\omega, t_{(j^*-1)}) < \infty$. It follows that

$$\sup_{p^*} \sum_{i=1}^{t_{j^*}-1} r(p^*, p_{\omega'}, i) \leq \left(\max_{j\leq j^*} \frac{d_j}{d_{j-1}} \right) \sup_{p^*} \sum_{i=1}^{t_{(j^*-1)}} r(p^*, p_{\omega''}, i) < \infty,$$

and similarly

$$\limsup_{n\to\infty} \frac{\sup_{p^*} \sum_{i=t_{j^*}}^{n} r(p^*, p_{\omega'}, i)}{\sup_{p^*} \sum_{i=t_{(j^*-1)}+1}^{n} r(p^*, p_{\omega''}, i)}$$
$$\leq \limsup_{n\to\infty} \sup_{p^*} \frac{\sum_{i=t_{j^*}}^{n} r(p^*, p_{\omega'}, i)}{\sum_{i=t_{(j^*-1)}+1}^{n} r(p^*, p_{\omega''}, i)} \leq \sup_{j\geq j^*} \frac{d_j}{d_{j-1}} \leq c + \varepsilon.$$

Consequently, using that $(\log n)^{2+\alpha}/R_{\mathrm{m}}(p_\omega, n) \to 0$ implies that $R_{\mathrm{m}}(p_\omega, n) \to \infty$, we find that

$$\begin{aligned}\limsup_{n\to\infty} \frac{R_{\mathrm{m}}(p_{\omega'}, n)}{R_{\mathrm{m}}(p_\omega, n)} &\leq \limsup_{n\to\infty} \frac{\sup_{p^*} \sum_{i=1}^{t_{j^*}-1} r(p^*, p_{\omega'}, i)}{R_{\mathrm{m}}(p_\omega, n)} \\ &\quad + \limsup_{n\to\infty} \frac{\sup_{p^*} \sum_{i=t_{j^*}}^{n} r(p^*, p_{\omega'}, i)}{\sup_{p^*} \sum_{i=t_{(j^*-1)}+1}^{n} r(p^*, p_{\omega''}, i)} \\ &\leq 0 + (c + \varepsilon),\end{aligned}$$

and (2.26) follows by letting ε tend to 0. □

2.9.3 Propositions 2.2 and 2.3

Both Proposition 2.2 and Proposition 2.3 follow from the following more general proposition.

Proposition 2.4. *Suppose that $\mathcal{M}^*$ is standard IID and p is an estimator such that $R_{\mathrm{m}}(p, n) \preceq c_3\, R_{\mathrm{mm}}(n)$ for some constant c_3. Define the time average (or Cesàro average)*

$$\bar{p}(X_n = x \mid x^{n-1}) = \frac{1}{n} \sum_{i=1}^{n} p(X_i = x \mid x^{i-1}).$$

Then

$$r_{\mathrm{mm}}(n) \leq r_{\mathrm{m}}(\bar{p}, n) \preceq c_3\, n^{-1} R_{\mathrm{mm}}(n) \leq c_3\, n^{-1} \sum_{i=1}^{n} r_{\mathrm{mm}}(i).$$

Furthermore, if $\mathcal{M}^*$ satisfies Condition 2.2 with c_1, c_2, γ and h_0 as in (2.21), and $r_{\mathrm{mm}}(n) < \infty$ for all n, then also

$$c_3\, n^{-1} \sum_{i=1}^{n} r_{\mathrm{mm}}(i) \preceq \frac{c_2}{c_1} \frac{c_3}{1-\gamma}\, r_{\mathrm{mm}}(n).$$

To obtain Proposition 2.3, let p be p_{sw}. To prove Proposition 2.2, note that by definition for every $\varepsilon > 0$ there exists an estimator p that achieves the minimax cumulative rate up to a factor $(1+\varepsilon)$, i.e. $R_{\mathrm{m}}(p, n) \preceq (1+\varepsilon) R_{\mathrm{mm}}(n)$. The proposition follows from Proposition 2.4 by letting ε tend to 0, such that c_3 tends to 1.

The proof of Proposition 2.4 requires the following lemma:

Lemma 2.2. *Let $g, h\colon \mathbb{Z}^+ \to \mathbb{R} \cup \{\infty\}$ be nonnegative functions such that $\sum_{i=1}^n h(i) \to \infty$ as n grows, and $g(i) < \infty$ for all i. Then $g(i) \preceq h(i)$ implies $\sum_{i=1}^n g(i) \preceq \sum_{i=1}^n h(i)$.*

Proof. Let $\varepsilon > 0$ be arbitrary. Then there exists an n_ε such that $g(i) \leq (1+\varepsilon)h(i)$ for all $i \geq n_\varepsilon$. Hence

$$\begin{aligned}\limsup_{n\to\infty} \frac{\sum_{i=1}^n g(i)}{\sum_{i=1}^n h(i)} &= \limsup_{n\to\infty} \frac{\sum_{i=1}^{n_\varepsilon - 1} g(i)}{\sum_{i=1}^n h(i)} + \limsup_{n\to\infty} \frac{\sum_{i=n_\varepsilon}^n g(i)}{\sum_{i=1}^n h(i)} \\ &\leq 0 + (1+\varepsilon).\end{aligned}$$

The lemma follows by letting ε tend to 0. □

Proof of Proposition 2.4. We show this by extending an argument from [Yang and Barron, 1999, p. 1582]. By applying Jensen's inequality as in Proposition 15.2 of [Grünwald, 2007] (or the corresponding results in [Yang, 2000] or [Yang and Barron, 1999]) it follows that, for all $p^* \in \mathcal{M}^*$, $r(p^*, \bar{p}, n) \leq \frac{1}{n} R(p^*, p, n)$, so that also

$$r_{\mathrm{m}}(\bar{p}, n) \leq \frac{1}{n} R_{\mathrm{m}}(p, n).$$

This implies that

$$n r_{\mathrm{mm}}(n) \leq n\, r_{\mathrm{m}}(\bar{p}, n) \leq R_{\mathrm{m}}(p, n) \preceq c_3\, R_{\mathrm{mm}}(n) \leq c_3 \sum_{i=1}^n r_{\mathrm{mm}}(i).$$

If $\mathcal{M}^*$ satisfies Condition 2.2, we further have:

$$\begin{aligned}\sum_{i=1}^n r_{\mathrm{mm}}(i) &\preceq c_2 \sum_{i=1}^n i^{-\gamma} h_0(i) \leq c_2\, h_0(n) \sum_{i=1}^n i^{-\gamma} \\ &\overset{(a)}{\leq} c_2 \frac{1}{1-\gamma} h_0(n) n^{1-\gamma} \preceq \frac{c_2}{c_1} \frac{1}{1-\gamma} n\, r_{\mathrm{mm}}(n),\end{aligned}$$

where the first step uses Lemma 2.2 and (a) follows by approximating the sum by an integral. The result follows. □

2.9.4 Proof of Theorem 2.2

The proof of Theorem 2.2 is based on the following lemma.

Lemma 2.3 (Fast Switching Lemma)**.** *Let $\mathcal{M}^*$ be standard IID and assume Condition 2.2 holds, with $c_1 n^{-\gamma} h_0(n) \preceq r_{\mathrm{mm}}(n) \preceq c_2 n^{-\gamma} h_0(n)$, as in (2.21). Suppose there exists an oracle ω relative to $\mathcal{L}_1, \mathcal{L}_2, \ldots$, with $r_{\mathrm{m}}(p_\omega, n) < \infty$ for all n, that achieves the minimax risk up to some nondecreasing function $f : \mathbb{Z}^+ \to [1, \infty)$, i.e. $r_{\mathrm{m}}(p_\omega, n) \preceq f(n)\, r_{\mathrm{mm}}(n)$. Let $\mathbf{t} = t_1, t_2, \ldots$ be the freezing times used to define $\mathcal{K}_1, \mathcal{K}_2, \ldots$ Then there exists an oracle ω' relative to $\mathcal{K}_1, \mathcal{K}_2, \ldots$ that switches only at times $\mathbf{t}$ and satisfies*

$$R_{\mathrm{m}}(p_{\omega'}, n) \preceq c\, f(n)\, R_{\mathrm{mm}}(n),$$

where c is as in (2.22).

The proof of Lemma 2.3 requires the following lemma:

Lemma 2.4. *If $\mathcal{M}^*$ is a standard IID model class that satisfies Condition 2.2, then $\sum_{i=1}^{n} r_{\mathrm{mm}}(i) \to \infty$ and for any sequence $1 = t_1 < t_2 < \cdots$ also $\sum_{j=1}^{m} d_j r_{\mathrm{mm}}(t_j) \to \infty$ (as a function of m), where $d_j = t_{j+1} - t_j$.*

Proof. Let $c_1 > 0$ and $0 \le \gamma < 1$ be constants and h_0 a nondecreasing, strictly positive function that satisfy Condition 2.2. Then by assumption there exists an n^* such that $r_{\mathrm{mm}}(i) \ge \frac{1}{2} c_1 i^{-\gamma} h_0(i)$ for all $i \ge n^*$. Hence

$$\sum_{i=1}^{n} r_{\mathrm{mm}}(i) \ge \sum_{i=n^*}^{n} r_{\mathrm{mm}}(i) \ge \sum_{i=n^*}^{n} c_1 h_0(i) i^{-\gamma} \ge c_1 h_0(1) \sum_{i=n^*}^{n} i^{-\gamma} \to \infty,$$

as required. Similarly, let j^* be sufficiently large that $t_{j^*} \ge n^*$. Then

$$\sum_{j=1}^{m} d_j r_{\mathrm{mm}}(t_j) \ge \sum_{j=j^*}^{m} d_j r_{\mathrm{mm}}(t_j) \ge c_1 h_0(1) \sum_{j=j^*}^{m} d_j t_j^{-\gamma}. \tag{2.27}$$

As $t_j^{-\gamma}$ is decreasing in t_j,

$$\sum_{j=j^*}^{m} d_j t_j^{-\gamma} \ge \sum_{i=t_{j^*}}^{t_{m+1}-1} i^{-\gamma} \to \infty.$$

Combining with (2.27) completes the proof. □

Proof of Lemma 2.3. Let $s(n)$ denote the last freezing time preceding n, i.e. $s(n) = t_k$ for k such that $t_k \le n < t_{k+1}$. Then for any p^*, n and x^{n-1},

ω' is defined such that it copies the prediction made by ω at time $s(n)$. That is,

$$\omega'(p^*, x^{n-1}) = \omega(p^*, x^{t_{s(n)}-1}) \circ s(n).$$

Thus, at any freezing time t_j, the predictions of ω and ω' coincide and $r(p^*, p_{\omega'}, t_j) = r(p^*, p_\omega, t_j)$.

Let us consider the blocks of indices between subsequent freezing times. For brevity, let $e_j = \min\{n, t_{j+1} - 1\}$ be the last index in block j and let $d_j = e_j - t_j + 1$ be the length of block j. For $m(n)$ such that $t_{m(n)} \leq n < t_{m(n)+1}$, we then have

$$\begin{aligned} R_{\mathrm{m}}(p_{\omega'}, n) &= \sup_{p^* \in \mathcal{M}^*} \sum_{i=1}^{n} r(p^*, p_{\omega'}, i) \leq \sum_{i=1}^{n} r_{\mathrm{m}}(p_{\omega'}, i) \\ &= \sum_{j=1}^{m(n)} d_j r_{\mathrm{m}}(p_{\omega'}, t_j) = \sum_{j=1}^{m(n)} d_j r_{\mathrm{m}}(p_\omega, t_j). \end{aligned}$$

As $f(t_j) \geq 1$, Lemma 2.4 implies that $\sum_{j=1}^{m} d_j f(t_j) r_{\mathrm{mm}}(t_j) \to \infty$. Therefore by Lemma 2.2

$$\sum_{j=1}^{m(n)} d_j r_{\mathrm{m}}(p_\omega, t_j) \preceq \sum_{j=1}^{m(n)} d_j f(t_j) r_{\mathrm{mm}}(t_j) \leq f(n) \sum_{j=1}^{m(n)} d_j r_{\mathrm{mm}}(t_j).$$

If $R_{\mathrm{mm}}(n)$ is infinite from some n onwards, then the lemma is trivially true. So assume that $R_{\mathrm{mm}}(n) < \infty$ for all n, which implies that $r_{\mathrm{mm}}(t_j) \leq R_{\mathrm{mm}}(t_j) < \infty$ for all t_j. Hence, again by Lemma 2.2 and using that h_0 is nondecreasing,

$$\begin{aligned} \sum_{j=1}^{m(n)} d_j r_{\mathrm{mm}}(t_j) &\preceq c_2 \sum_{j=1}^{m(n)} d_j t_j^{-\gamma} h_0(t_j) \leq c_2 \sum_{j=1}^{m(n)} \sum_{i=t_j}^{e_j} \left(\frac{i}{t_j}\right)^{\gamma} i^{-\gamma} h_0(i) \\ &\leq c_2 \sup_{j \geq 1} \left\{ \left(\frac{t_{j+1} - 1}{t_j} \right)^{\gamma} \right\} \sum_{i=1}^{n} i^{-\gamma} h_0(i). \end{aligned}$$

By Lemma 2.4, $\sum_{i=1}^{n} r_{\mathrm{mm}}(i) \to \infty$. Therefore by Lemma 2.2

$$\sum_{i=1}^{n} i^{-\gamma} h_0(i) \preceq \frac{1}{c_1} \sum_{i=1}^{n} r_{\mathrm{mm}}(i).$$

Finally, by Proposition 2.2

$$\sum_{i=1}^{n} r_{\mathrm{mm}}(i) \preceq \frac{c_2}{c_1} \frac{1}{1-\gamma} R_{\mathrm{mm}}(n).$$

The result is obtained by combining all the bounds above. □

Proof of Theorem 2.2. By Lemma 2.3 there exists an oracle ω' relative to $\mathcal{K}_1, \mathcal{K}_2, \ldots$ that switches only at times $\mathbf{t}$ and is such that

$$R_{\mathrm{m}}(p_{\omega'}, n) \preceq c f(n) R_{\mathrm{mm}}(n). \tag{2.28}$$

Let $m(n)$ denote the maximum number of different prediction strategies ω' uses before the n-th outcome, as in (2.25). Then the choice of $\mathbf{t}$ ensures that $m(n) = O(\log n)$, such that by the Oracle Approximation Lemma (Lemma 2.1) and Condition 2.1

$$R_{\mathrm{m}}(p_{\mathrm{sw}}, n) = R_{\mathrm{m}}(p_{\omega'}, n) + O\left((\log n)^2\right). \tag{2.29}$$

Finally, Proposition 2.2 and Condition 2.2 together imply that $R_{\mathrm{mm}}(n) \succeq n r_{\mathrm{mm}}(n) \succeq c_1 h_0(1) n^{1-\gamma}$, so that $(\log n)^2 / R_{\mathrm{mm}}(n) \to 0$. Combining this with (2.28) and (2.29), the result follows. □

2.10 Consistency Proof

This section gives the proof of Theorem 2.3 from Section 2.6.

2.10.1 Proof of Theorem 2.3

Proof. It is sufficient to show that

$$\lim_{n\to\infty} p_{\mathrm{sw}}(K_{n+1} \neq k^* \mid X^n) = 0 \qquad (p_{k^*}\text{-a.s.}), \tag{2.30}$$

which is equivalent to (2.24) except that p_{θ^*}-probability has been replaced by p_{k^*}-probability. To see this, suppose the theorem is false. Then there exists a set of parameters $\Phi \subseteq \Theta_{k^*}$ with $w_{k^*}(\Phi) > 0$ such that (2.24) does not hold for any $\theta^* \in \Phi$. But then by definition of p_{k^*}, which is a mixture of p_θ with weights $w(\theta)$, we have a contradiction with (2.30).

For any n, let $U_n = \{\mathbf{s} \in \mathbb{S} \mid K_{n+1}(\mathbf{s}) \neq k^*\}$ denote the set of "bad" parameters that select an incorrect model. Let n' be the smallest $n \geq n_{k^*}$ such that $|\text{support}(\lambda_{n+1})| > 1$. (Note that $n' > n_{k^*}$ only in the degenerate case that $\lambda_n(k^*) = 1$ for all $n \leq n'$.) The assumption that $\sum_{\mathbf{s} \in B^{k^*}_{n_{k^*}}} \pi(\mathbf{s}) q_{\mathbf{s}}(X^{n_{k^*}}) > 0$ (p_{k^*}-a.s.) implies that

$$p_{\text{sw}}(X^{n'}) \geq \sum_{\mathbf{s} \in B^{k^*}_{n_{k^*}}} \pi(\mathbf{s}) q_{\mathbf{s}}(X^{n_{k^*}}) p_{k^*}(X^{n'}_{n_{k^*}+1} \mid X^{n_{k^*}}) > 0 \qquad (p_{k^*}\text{-a.s.}),$$

where $X^b_a = X_a, \ldots, X^b$. Hence the posterior distribution

$$\pi(\mathbf{s} \mid X^{n'}) = \frac{\pi(\mathbf{s}) q_{\mathbf{s}}(X^{n'})}{p_{\text{sw}}(X^{n'})}$$

is defined (p_{k^*}-a.s.), and by substituting definitions we find that (2.30) is equivalent to

$$\lim_{n \to \infty} \frac{p_{\text{sw}}(X^{n'}) \sum_{\mathbf{s} \in U_n} \pi(\mathbf{s} \mid X^{n'}) q_{\mathbf{s}}(X^n_{n'+1} \mid X^{n'})}{p_{\text{sw}}(X^n)} = 0 \qquad (p_{k^*}\text{-a.s.}). \quad (2.31)$$

There are two reasons why a parameter $\mathbf{s} = ((t_1, k_1), \ldots, (t_m, k_m))$ may be in U_n: either $t_m(\mathbf{s}) \leq n+1$ and $k_m \neq k^*$ or $t_m > n+1$ and $K_{n+1}(\mathbf{s}) \neq k^*$. Note that the second case may occur even when the final prediction strategy k_m equals k^*. We would like to get rid of such parameters and replace U_n by the set

$$A = \{\mathbf{s} = \Big((t_1, k_1), \ldots, (t_m, k_m)\Big) \in \mathbb{S} \mid k_m \neq k^*, \pi(\mathbf{s}) > 0\},$$

which does not depend on n. To this end, fix any $k' \neq k^*$ with $\lambda_{n'+1}(k') > 0$. We define an alternative distribution $\pi'(\mathbf{s} \mid X^{n'})$, which is equal to $\pi(\mathbf{s} \mid X^{n'})$, except that it puts all probability mass from any parameter such that $k_m = k^*$ on a corresponding parameter, which is identical except that $k_m = k'$. That is,

$$\pi'\Big(((t_1, k_1), \ldots, (t_m, k_m)) \mid X^{n'}\Big) = \begin{cases} 0 & \text{if } k_m = k^*; \\ \sum_{k \in \{k^*, k'\}} \pi\Big(((t_1, k_1), \ldots, (t_m, k)) \mid X^{n'}\Big) & \text{if } k_m = k'; \\ \pi\Big(((t_1, k_1), \ldots, (t_m, k_m)) \mid X^{n'}\Big) & \text{otherwise.} \end{cases}$$

Suppose $\mathbf{s} = ((t_1, k_1), \ldots, (t_m, k^*))$ is a parameter with $k_m = k^*$ and $\mathbf{s}' = ((t_1, k_1), \ldots, (t_m, k'))$ is the corresponding parameter with $k_m = k'$. Then if $t_m > n+1$, we have that $q_{\mathbf{s}}(X_{n'+1}^n \mid X^{n'}) = q_{\mathbf{s}'}(X_{n'+1}^n \mid X^{n'})$; and if $t_m \leq n+1$, then $\mathbf{s} \notin U_n$. It follows that

$$\sum_{\mathbf{s} \in U_n} \pi(\mathbf{s} \mid X^{n'}) q_{\mathbf{s}}(X_{n'+1}^n \mid X^{n'}) \leq \sum_{\mathbf{s} \in A} \pi'(\mathbf{s} \mid X^{n'}) q_{\mathbf{s}}(X_{n'+1}^n \mid X^{n'}),$$

which gives a bound on the numerator of (2.31). We may also bound the denominator by

$$p_{\mathrm{sw}}(X^n) \geq \left(\sum_{\mathbf{s} \in B_{n_{k^*}}^{k^*}} \pi(\mathbf{s}) q_{\mathbf{s}}(X^{n_{k^*}}) \right) p_{k^*}(X_{n_{k^*}+1}^{n'} \mid X^{n_{k^*}}) p_{k^*}(X_{n'+1}^n \mid X^{n'}).$$

As $\left(\sum_{\mathbf{s} \in B_{n_{k^*}}^{k^*}} \pi(\mathbf{s}) q_{\mathbf{s}}(X^{n_{k^*}}) \right) p_{k^*}(X_{n_{k^*}+1}^{n'} \mid X^{n_{k^*}})$ is positive (p_{k^*}-a.s.), it is therefore sufficient to show that

$$\lim_{n \to \infty} \frac{r(X_{n'+1}^n \mid X^{n'})}{p_{k^*}(X_{n'+1}^n \mid X^{n'})} = 0 \qquad (p_{k^*}\text{-a.s.}), \tag{2.32}$$

where $r(X_{n'+1}^n \mid X^{n'}) = \sum_{\mathbf{s} \in A} \pi'(\mathbf{s} \mid X^{n'}) q_{\mathbf{s}}(X_{n'+1}^n \mid X^{n'})$ is a countable mixture of prediction strategies $q_{\mathbf{s}}$ that eventually switch to a prediction strategy p_{k_m} that is mutually singular with p_{k^*} by assumption.

Suppose first that $n' = 0$. Then suppose $\mathbf{s} = ((t_1, k_1), \ldots, (t_m, k_m)) \in A$. (Note that this implies that $\lambda_{t_m}(k_m) > 0$.) It can be shown that mutual singularity of $p_{k_m}(X_{t_m}^\infty \mid X^{t_m - 1})$ and $p_{k^*}(X_{t_m}^\infty \mid X^{t_m - 1})$ (p_{k^*}-a.s.), which we have assumed, implies mutual singularity of $q_{\mathbf{s}}(X^\infty)$ and $p_{k^*}(X^\infty)$ (p_{k^*}-a.s.) To see this for countable $\mathcal{X}$, let $E_{X^{t_m - 1}} \subseteq \mathcal{X}_{t_m} \times \mathcal{X}_{t_m+1} \times \cdots$ be an event such that $p_{k_m}(E_{X^{t_m-1}} \mid X^{t_m-1}) = 1$ and $p_{k^*}(E_{X^{t_m-1}} \mid X^{t_m-1}) = 0$. Then, for $E = \{X^\infty \in \mathcal{X}^\infty \mid X_{t_m}^\infty \in E_{X^{t_m}}\}$, we have that $q_{\mathbf{s}}(E) = 1$ and $p_{k^*}(E) = 0$. In the uncountable case, however, the set E may not be measurable. In that case, mutual singularity follows by Corollary 2.3 proved below, which only relies on the fact that $\mathcal{X} \subseteq \mathbb{R}^d$ is a separable metric space.

As $r(X^\infty)$ is a countable mixture of distributions that are mutually singular with $p_{k^*}(X^\infty)$, it is itself mutually singular with $p_{k^*}(X^\infty)$. This implies (2.32), because the density ratio $r(X^n)/p_{k^*}(X^n)$ tends to $r(X^\infty)/p_{k^*}(X^\infty)$ with p_{k^*}-probability 1 (e.g. by Lévy's theorem

[Shiryaev, 1996]), which is zero with probability 1 by mutual singularity of $r(X^\infty)$ and $p_{k^*}(X^\infty)$.

It remains to show (2.32) when $n' > 0$. In this case it is seen that all properties that were required for the case $n' = 0$ continue to hold with p_{k^*}-probability 1 when all distributions are conditioned on $X^{n'}$. This completes the proof. □

2.10.2 Mutual Singularity as Used in the Proof of Theorem 2.3

Let $Y^2 = (Y_1, Y_2)$ be random variables that take values in separable metric spaces Ω_1 and Ω_2, respectively. We will assume all spaces to be equipped with Borel σ-algebras generated by the open sets. Let p and q be prediction strategies for Y^2.

Lemma 2.5. *If $p(Y_2 \mid Y_1)$ and $q(Y_2 \mid Y_1)$ are mutually singular (p-a.s.), then $p(Y^2)$ and $q(Y^2)$ are mutually singular.*

The proof is given below the following corollary, which is what we are really interested in. Let $X^\infty = X_1, X_2, \ldots$ be random variables that take values in the separable metric space $\mathcal{X}$. Then what we need in the proof of Theorem 2.3 is the following corollary of Lemma 2.5:

Corollary 2.3. *Suppose p and q are prediction strategies for X^∞ and let n be any positive integer. If $p(X^\infty \mid X^n)$ and $q(X^\infty \mid X^n)$ are mutually singular (p-a.s.), then $p(X^\infty)$ and $q(X^\infty)$ are mutually singular.*

Proof. The product spaces $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ and $\mathcal{X}_{n+1} \times \mathcal{X}_{n+2} \times \cdots$ are separable metric spaces [Parthasarathy, 1967, pp. 5,6]. Now apply Lemma 2.5 with $\Omega_1 = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ and $\Omega_2 = \mathcal{X}_{n+1} \times \mathcal{X}_{n+2} \times \cdots$. □

Proof of Lemma 2.5. Let $\Gamma \subseteq \Omega_1$ be a measurable set such that $p(\Gamma) = 1$ and $p(Y_2 \mid \omega_1)$ and $q(Y_2 \mid \omega_1)$ are mutually singular for all $\omega_1 \in \Gamma$. Then for each $\omega_1 \in \Gamma$ there exists a measurable set $C_{\omega_1} \subseteq \Omega_2$ such that $p(C_{\omega_1} \mid \omega_1) = 1$ and $q(C_{\omega_1} \mid \omega_1) = 0$. As Ω_2 is a metric space, it follows from [Parthasarathy, 1967, Theorems 1.1 and 1.2 in Chapter II] that for any $\epsilon > 0$ there exists an open set $U^\epsilon_{\omega_1} \supseteq C_{\omega_1}$ such that

$$p(U^\epsilon_{\omega_1} \mid \omega_1) = 1 \quad \text{and} \quad q(U^\epsilon_{\omega_1} \mid \omega_1) < \epsilon. \tag{2.33}$$

As Ω_2 is a separable metric space, there also exists a countable sequence $\{B_i\}_{i \geq 1}$ of open sets such that every open subset of Ω_2

($U^\epsilon_{\omega_1}$ in particular) can be expressed as the union of sets from $\{B_i\}$ [Parthasarathy, 1967, Theorem 1.8 in Chapter I].

Let $\{B'_i\}_{i\geq 1}$ denote a subsequence of $\{B_i\}$ such that $U^\epsilon_{\omega_1} = \bigcup_i B'_i$. Suppose $\{B'_i\}$ is a finite sequence. Then let $V^\epsilon_{\omega_1} = U^\epsilon_{\omega_1}$. Suppose it is not. Then $1 = p(U^\epsilon_{\omega_1} \mid \omega_1) = p(\bigcup_{i=1}^\infty B'_i \mid \omega_1) = \lim_{n\to\infty} p(\bigcup_{i=1}^n B'_i \mid \omega_1)$, because $\bigcup_{i=1}^n B'_i$ as a function of n is an increasing sequence of sets. Consequently, there exists an N such that $p(\bigcup_{i=1}^N B'_i \mid \omega_1) > 1-\epsilon$ and we let $V^\epsilon_{\omega_1} = \bigcup_{i=1}^N B'_i$. Thus in any case there exists a set $V^\epsilon_{\omega_1} \subseteq U^\epsilon_{\omega_1}$ that is a union of a finite number of elements in $\{B_i\}$ such that

$$p(V^\epsilon_{\omega_1} \mid \omega_1) > 1-\epsilon \quad \text{and} \quad q(V^\epsilon_{\omega_1} \mid \omega_1) < \epsilon. \tag{2.34}$$

Let $\{D\}_{i\geq 1}$ denote an enumeration of all possible unions of a finite number of elements in $\{B_i\}$ and define the disjoint sequence of sets $\{A^\epsilon_i\}_{i\geq 1}$ by

$$A^\epsilon_i = \{\omega_1 \in \Gamma\colon p(D_i \mid \omega_1) > 1-\epsilon, q(D_i \mid \omega_1) < \epsilon\} \setminus \bigcup_{j=1}^{i-1} A^\epsilon_j \tag{2.35}$$

for $i = 1, 2, \ldots$ Note that, by the reasoning above, for each $\omega_1 \in \Gamma$ there exists an i such that $\omega_1 \in A^\epsilon_i$, which implies that $\{A^\epsilon_i\}$ forms a partition of Γ. Now, as all elements of $\{A^\epsilon_i\}$ and $\{D_i\}$ are measurable, so is the set $F^\epsilon = \bigcup_{i=1}^\infty A^\epsilon_i \times D_i \subseteq \Omega_1 \times \Omega_2$, for which we have that $p(F^\epsilon) = \sum_{i=1}^\infty p(A^\epsilon_i \times D_i) > (1-\epsilon)\sum_{i=1}^\infty p(A^\epsilon_i) = 1-\epsilon$ and likewise $q(F^\epsilon) < \epsilon$.

Finally, let $G = \bigcap_{n=1}^\infty \bigcup_{k=n}^\infty F^{2^{-k}}$. Then

$$p(G) = \lim_{n\to\infty} p(\textstyle\bigcup_{k=n}^\infty F^{2^{-k}}) \geq \lim_{n\to\infty} 1 - 2^{-n} = 1$$

and

$$q(G) = \lim_{n\to\infty} q(\textstyle\bigcup_{k=n} F^{2^{-k}}) \leq \lim_{n\to\infty} \displaystyle\sum_{k=n}^\infty 2^{-k} = \lim_{n\to\infty} 2^{-n+1} = 0, \tag{2.36}$$

which proves the lemma. □

From Prediction Strategies to Experts

The next two chapters provide a further study of switching between prediction strategies. They consider extensions and variations of the `Fixed-share` algorithm, which is also generalised by the switch distribution as discussed on page 60. We will adopt the notation and conventions of the related literature on *prediction with expert advice*, in which prediction strategies are replaced by the more general notion of *experts*. To relate the results to the previous chapter, one may think of an expert as a "meta" prediction strategy relative to some parametric model, as described in Section 2.3.1. The main change in notation is that n denotes the number of experts in Chapter 3, instead of the sample size.

The setup of prediction with expert advice is online: prediction proceeds in rounds $1, 2, \ldots$ and no fixed sample size is assumed to be known in advance. Instead, guarantees about cumulative prediction error need to hold simultaneously for any number of outcomes T (although the guarantees will typically depend on T).

The main difference with the previous chapter is that instead of comparing predictive performance to an assumed true distribution p^* in expectation under p^*, we will now compare ourselves to the best expert for the data on the worst-case data. The guarantees we obtain therefore do not depend on any probabilistic assumptions, not even that the data are i.i.d. They do depend on the quality of the best expert's predictions however: our bounds are only useful if the best expert actually manages to predict the data well.

Overview The `Fixed-share` algorithm takes a parameter called the *switching rate*, which determines how much prior weight it assigns to switching between prediction strategies. In Chapter 3 a fully online algorithm is presented that learns the optimal switching rate, and its running time and predictive performance are compared to previous approaches to learning the switching rate.

In Chapter 4 the influence of switches on the experts' predictions is considered. One may distinguish between two cases:

1. Firstly, the reason to switch between experts may be that the (relative) quality of the experts' predictions changes over time, although the nature of the data stays the same. The experts are assumed not to care about the timing of switches. This may be the case, for example, if the data are generated by an i.i.d. process and the experts are based on estimators that predict better the more data they have seen, like in Chapter 2.

2. Alternatively, it may the case that the nature of the data does change over time. For example, one may think of weather data that depend on the season. Then the reason to switch between experts may be that different experts are good for different periods (e.g. seasons). In this case the experts themselves should also care about the timing of the switches, because learning from data from the wrong period (season) will throw off their predictions.

The switch distribution has been designed for case 1, and this is also the case for which the `Fixed-Share` algorithm is appropriate. Paradoxically, however, the `Fixed-Share` algorithm has been designed with case 2 in mind. In Chapter 4 we therefore consider how to modify `Fixed-Share` for case 2. It is shown that if the expert predictions have internal structure that can be represented by so-called *expert hidden Markov models*, then the modified `Fixed-share` algorithm can automatically and efficiently feed them only data from the appropriate period to learn from.

Chapter 3

Learning the Switching Rate by Discretising Bernoulli Sources Online

The expert tracking algorithm `Fixed-share` depends on a parameter α, called the *switching rate*. The switching rate can be learned online with regret $\frac{1}{2}\log T + O(1)$ bits. The current fastest method to achieve this is based on optimal discretisation of the Bernoulli distributions into $O(\sqrt{T})$ bins and runs in $O(T\sqrt{T})$ time. However, the exact locations of these bins have to be determined algorithmically, and the final number of outcomes T must be known in advance.

This chapter introduces a new discretisation scheme with the same regret bound for known T, that specifies the number and positions of the discretisation points explicitly. The scheme is especially useful, however, when T is not known in advance: a new fully online algorithm is presented, which runs in $O(T\sqrt{T}\log T)$ time and achieves a regret of $\frac{1}{2}\log 3 \log T + O(\log\log T)$ bits.

3.1 Introduction

We will attempt to sequentially predict the outcomes X_1, X_2, ... from an unknown process, where each outcome takes values in a countable set $\mathcal{X}$. At each time $t \in \mathbb{Z}^+ = \{1, 2, \ldots\}$ we have to issue a probability distribution $P(X_t \mid x^{t-1})$ on $\mathcal{X}$, which is allowed to depend on past observations $x^{t-1} = x_1, \ldots, x_{t-1}$. Then x_t is revealed and we suffer *logarithmic loss* $-\ln P(X_t = x_t \mid x^{t-1})$. (For simplicity we consider only logarithmic loss, but results for other loss functions can be obtained using methods described in e.g. [Vovk, 1999].) Suppose our understanding of the process is very limited, but luckily we do have access to n *experts*. Each expert $\xi \in \Xi = \{1, \ldots, n\}$ provides us with her prediction

$P_\xi(X_t \mid x^{t-1})$, on which we may base our own forecast $P(X_t \mid x^{t-1})$. We make no assumptions about the nature of the experts, so one may think of human experts, but also of computer algorithms. This is the problem of *prediction with expert advice* (for log loss) [Cesa-Bianchi and Lugosi, 2006].

For any T, one may view the predictions $P(X_t \mid X^{t-1})$ as conditionals of the joint distribution $P(X^T) = \prod_{t=1}^T P(X_t \mid X^{t-1})$. (We regard the empty sequence x^0 as a certain event, which occurs with probability one.) In its most basic setup the goal of prediction with expert advice is to minimise the excess loss compared to the best expert on any sequence of outcomes x^T:

$$-\ln P(x^T) - \min_\xi [-\ln P_\xi(x^T)].$$

This is called the *regret* on x^T. A more ambitious goal is to compare to the performance that can be obtained by optimally dividing the data into m segments and, within each segment, using the best expert for that segment. This is prudent in case the experts themselves may improve (study hard) or deteriorate (take to drinking), but also when their performance depends on the predictive context (some experts may be good during spring, others during winter). In this case, if the optimal segments start at times $t_1, \ldots, t_m$ for a given sequence x^T, the goal is to minimise

$$-\ln P(x^T) - \sum_{i=1}^m \min_\xi -\ln P_\xi(x_{t_i}^{t_{i+1}-1} \mid x^{t_i-1}), \tag{3.1}$$

where $x_a^b = x_a, \ldots, x_b$, and $t_{m+1} = T+1$. This is the approach taken by Herbster and Warmuth [1998]; see also [Vovk, 1999, Cesa-Bianchi and Lugosi, 2006].

Let $H(p) = -p\ln p - (1-p)\ln(1-p)$ and $D(p\|q) = p\ln p/q + (1-p)\ln(1-p)/(1-q)$ denote the entropy and Kullback-Leibler divergence for a binary space, respectively; in this chapter, we use ln to denote the natural logarithm and log for base two. The regret of Herbster and Warmuth's `Fixed-share` algorithm is bounded from above by

$$(T-1)\Big(H(\alpha^*) + D(\alpha^*\|\alpha)\Big) + (m-1)\ln(n-1) + \ln n$$

nats (see Theorem 4.1 in Chapter 4), where $\alpha^* := (m-1)/(T-1)$ and α is the *switching rate*, a parameter of the algorithm that can be interpreted

as the probability of switching between experts. In Figure 2.2 from Chapter 2 this parameter α was called $\tau_t = \tau(Z = t+1 \mid Z > t)$. The best regret bound is obtained when α equals α^*.

One clear advantage of `Fixed-share` is its computational efficiency: its running time, which is $n \cdot O(T)$, is as low as that of the standard Bayesian mixture. The one real disadvantage is having to specify the switching rate. It is this problem that we address in this chapter. Our contribution should be placed in the context of three earlier approaches to avoid a priori specification of the switching rate:

Decreasing Switching Rate One option is to let the switching rate *decrease with time* as $1/t$. This corresponds to using $\tau(t) = 1/(t(t-1))$, for which $\tau(Z = t \mid Z > t-1) = 1/t$ (see also [Koolen and de Rooij, 2008a]). For this approach, the regret compared to the best segmentation in m parts is within $\ln T + O(m \log m)$ nats from the bound for `Fixed-share` with optimally tuned α. This is fine if the number of switches in the sequence is not too large (say, $m = O(\log T)$), but if switches can occur more frequently, it may not be the best choice.

Bayes with Undiscretised Switching Rate A second option is to use a Bayesian mixture over α. Such an algorithm was described very early in the source coding literature [Volf and Willems, 1998]. This algorithm, called the *Switching Method* (not to be confused with the switch distribution!), achieves a regret bounded by $\frac{1}{2} \ln T + O(1)$ nats compared to the best `Fixed-share` parameter. Note that this bound does not depend on the number of switches. The drawback of this approach is that its running time is $n \cdot O(T^2)$, which is significantly slower than the previous algorithms and may be prohibitive in some applications.

Bayes with Discretised Switching Rate A third approach to get rid of α also uses a Bayesian mixture, but rather than putting a prior on the whole range $[0, 1]$ of possible values of α, a prior is defined on a *discretised* set of parameters $\alpha_1, \alpha_2, \ldots, \alpha_j$. Monteleoni and Jaakkola [2003] argue that $O(\sqrt{T})$ levels of discretisation suffice to achieve a regret with respect to `Fixed-share` of at most $\frac{1}{2} \ln T + O(1)$ nats, like the Switching Method. Their algorithm `Learn-`α has running time $n \cdot O(T\sqrt{T})$, a significant improvement over the Switching Method. However, while `Learn-`α does not require a priori knowledge of α, unlike the other

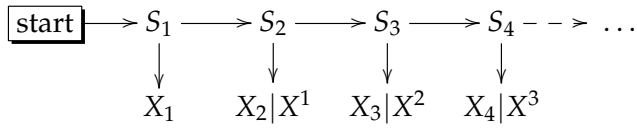


Figure 3.1: Bayesian network for an expert algorithm

approaches it does require a priori knowledge of the final number of outcomes T. The algorithm is therefore almost, but not completely, online. In Section 3.3.3 we discuss why the so-called doubling trick is not the best way to eliminate this dependence.

Refine-Online Here we take the `Learn-`α algorithm as a starting point to develop a fourth, fully online algorithm called `Refine-Online`. It has running time $n \cdot O(T\sqrt{T}\log T)$, which makes it only slightly slower than `Learn-`α. Its regret is bounded by $\frac{1}{2}\log 3 \ln T + \log 3 \ln \ln(T+1) + O(1)$, which is worse than the bounds in the two Bayesian approaches, but would still seem an acceptable price to pay to get a fast algorithm that is completely online.

Outline In Section 3.2 we show how probabilistic algorithms for prediction with expert advice can be described using Hidden Markov models (HMMs), and we give basic tools to prove loss bounds for such algorithms. We then state our main results. Section 3.3 exhibits a new, very simple discretisation scheme that grants full control over the exact number and placement of discretisation points, in contrast to the discretisation used by `Learn-`α, which can only be determined algorithmically. Moreover, we show how this discretisation can be refined online, so that the final number of outcomes T does not have to be known.

3.2 Expert Algorithms as HMMs

Many algorithms for prediction with expert advice can be described as a hidden Markov model (HMM) $\mathbb{P}$, where the *hidden state* S_t at any time t identifies an expert ξ_t to predict outcome X_t [Koolen and de Rooij, 2008a]. Figure 3.1 depicts the corresponding Bayesian network, where we write $X_t|X^{t-1}$ to emphasize that the expert may base her prediction

of X_t on all previous outcomes X^{t-1}. Each S_t takes values in a set of *hidden states* $\mathcal{S} = \{\langle \xi, t, \ldots \rangle \mid \xi \in \Xi, t \in \mathbb{Z}^+\}$, where t denotes a time index and states with the wrong time index get probability zero: $\mathbb{P}(S_t = \langle \xi, t', \ldots \rangle) = 0$ if $t' \neq t$. Depending on the specifics of the algorithm the hidden states can contain more information, represented here by dots. Given a state $\langle \xi, t, \ldots \rangle \in \mathcal{S}$ and previous outcomes x^{t-1} the probability of X_t is determined by the prediction of expert ξ:

$$\mathbb{P}(X_t \mid \langle \xi, t, \ldots \rangle, x^{t-1}) = P_\xi(X_t \mid x^{t-1}).$$

The advantage of casting these algorithms as HMMs is that the standard algorithms for HMMs can be applied. Specifically, the *forward algorithm* can compute the predictions $\mathbb{P}(X_1)$, ..., $\mathbb{P}(X_T \mid x^{T-1})$ in time proportional to the number of transitions in the HMM [Rabiner, 1989, Koolen and de Rooij, 2008a].

Bayes We first consider the standard Bayesian prediction strategy that puts a prior w on experts Ξ. This corresponds to the HMM $\mathbb{H}$ with hidden states $\{\langle \xi, t \rangle \mid \xi \in \Xi, t \in \mathbb{Z}^+\}$. Initially all experts get probability according to the prior, $\mathbb{H}(\langle \xi_1, 1 \rangle) = w(\xi_1)$, but afterwards no more switches between experts are allowed: $\mathbb{H}(\langle \xi_{t+1}, t+1 \rangle \mid \langle \xi_t, t \rangle)$ is 1 if $\xi_{t+1} = \xi_t$, and 0 otherwise.

Fixed-share There is also an HMM $\mathbb{F}_\alpha$ that corresponds to the `Fixed-share` algorithm [Koolen and de Rooij, 2008a]. As in [Herbster and Warmuth, 1998], all experts are initially given equal weight, $\mathbb{F}_\alpha(\langle \xi_1, 1 \rangle) = 1/n$, which gives the best worst-case bound. After each outcome, $\mathbb{F}_\alpha$ allows *switches* between experts to occur with probability $\alpha \in [0, 1]$, which is called the *switching rate*:

$$\mathbb{F}_\alpha(\langle \xi_{t+1}, t+1 \rangle \mid \langle \xi_t, t \rangle) = \begin{cases} 1 - \alpha & \text{if } \xi_{t+1} = \xi_t, \\ \alpha/(n-1) & \text{otherwise.} \end{cases}$$

Note that $\mathbb{F}_0 = \mathbb{H}$ (using a uniform prior w). Naive application of the forward algorithm to $\mathbb{F}_\alpha$ gives $O(n^2)$ transitions per time step, adding up to a total running time of $n^2 \cdot O(T)$. This is reduced to $O(n)$ transitions by introducing an intermediate *pool* state that first collects all probability mass for switches between experts and then redistributes it (see Figure 2.2 or [Koolen and de Rooij, 2008a] for details). The running time then becomes $n \cdot O(T)$ as in [Herbster and Warmuth, 1998].

3.2.1 Tracking HMMs and Bernoulli HMMs

The `Fixed-share` algorithm has a fixed switching rate α. This may be generalised to a *tracking HMM* $\mathbb{S}$ with hidden states $\{\langle \xi, t, \alpha \rangle \mid \xi \in \Xi, t \in \mathbb{Z}^+, \alpha \in \mathcal{A}_t\}$. The initial states have weights given by $\mathbb{S}(\langle \xi_1, 1, \alpha_1 \rangle) = \mathbb{B}(\langle \alpha_1, 1 \rangle) \cdot \frac{1}{n}$, and the transition probabilities are

$$\begin{aligned}&\mathbb{S}(\langle \xi_{t+1}, t+1, \alpha_{t+1} \rangle \mid \langle \xi_t, t, \alpha_t \rangle) \\ &= \mathbb{B}(\langle \alpha_{t+1}, t+1 \rangle \mid \langle \alpha_t, t \rangle) \cdot \mathbb{F}_{\alpha_t}(\langle \xi_{t+1}, t+1 \rangle \mid \langle \xi_t, t \rangle),\end{aligned}$$

where $\mathbb{B}$, called a *Bernoulli HMM*, describes the evolution of α. The original `Fixed-share` method $\mathbb{F}_\alpha$ can be recovered by using $\mathcal{A}_t = \{\alpha\}$ and $\mathbb{B} = \mathbb{B}^\alpha_{\text{fixed}}$, where

$$\mathbb{B}^\alpha_{\text{fixed}}(\langle \alpha_{t+1}, t+1 \rangle \mid \langle \alpha_t, t \rangle) = \mathbb{B}^\alpha_{\text{fixed}}(\langle \alpha_{t+1}, 1 \rangle) = 1.$$

We consider various other options for the Bernoulli HMM $\mathbb{B}$ as well. In general let $\mathbb{S}^b_a$ denote the tracking HMM $\mathbb{S}$ defined with respect to the Bernoulli HMM $\mathbb{B}^b_a$. Thus $\mathbb{S}^\alpha_{\text{fixed}} = \mathbb{F}_\alpha$.

It is essential now to distinguish between two levels: `Fixed-share` and the tracking HMM $\mathbb{S}$, which aim to predict outcomes $X_1, X_2, \ldots$, operate on the upper level. On the lower level there is the Bernoulli HMM $\mathbb{B}$. Although $\mathbb{B}$ is used as a building block in the construction of $\mathbb{S}$, it is convenient to also interpret $\mathbb{B}$ as an algorithm for prediction with expert advice in itself. In this view, let $Y_1, Y_2, \ldots$ be binary outcomes, which $\mathbb{B}$ has to predict, and let P_α denote the Bernoulli distribution with $P_\alpha(Y = 1) = \alpha$, extended to sequences by taking product distributions. In a Bernoulli HMM the experts are instantiated to such Bernoulli sources, and are indexed by $\alpha \in \mathcal{A}_t$. Thus $\mathbb{B}$ has hidden states $\{\langle \alpha, t \rangle \mid \alpha \in \mathcal{A}_t, t \in \mathbb{Z}^+\}$ and $\mathbb{B}(Y_t \mid \langle \alpha, t \rangle) = P_\alpha(Y_t)$.

The total running time of the forward algorithm applied to a tracking HMM may be computed by summing up the number of transitions for each time step. This is the number of transitions of `Fixed-share`, which is $O(n)$, times the number of transitions of the corresponding Bernoulli HMM. Thus the forward algorithm for a tracking HMM runs in $O(n)$ times the running time of the forward algorithm for its Bernoulli HMM.

All approaches to learning the switching rate that were discussed in the introduction, including the new `Refine-Online` method, can be implemented using tracking HMMs with different choices for the

Bernoulli HMM $\mathbb{B}$. We will illustrate this for `Learn-`α. In Section 3.3.3 we do the same for $\mathbb{B}_{\text{ro}}$, which defines the `Refine-Online` algorithm. From the description of the Switching Method in [Koolen and de Rooij, 2008a] it is not hard to see how it can be cast as a Bernoulli HMM as well, but for brevity we do not discuss the details here.

Example: `Learn-`α Given the final number of outcomes, T, the algorithm `Learn-`α [Monteleoni and Jaakkola, 2003] applies Bayes at a meta-level to learn the switching rate α of the `Fixed-share` algorithm: it puts a uniform prior (which gives the best worst-case bound) on a discretised set $\mathcal{A}_T$ of switching rates, where the discretisation depends on T. It turns out that this approach corresponds exactly to a tracking HMM $\mathbb{S}_{\text{Bayes}}$. The corresponding Bernoulli HMM $\mathbb{B}_{\text{Bayes}}$ has $\mathcal{A}_t = \mathcal{A}_T$ for all t, initial weights $\mathbb{B}_{\text{Bayes}}(\langle \alpha_1, 1\rangle) = 1/|\mathcal{A}_T|$ and transition probabilities

$$\mathbb{B}_{\text{Bayes}}(\langle \alpha_{t+1}, t+1\rangle \mid \langle \alpha_t, t\rangle) = \mathbf{1}_{\{\alpha_t\}}(\alpha_{t+1}), \tag{3.2}$$

where $\mathbf{1}_A(z)$ denotes the *indicator function*, which is 1 if $z \in A$ and 0 otherwise. Note that $\mathbb{B}_{\text{Bayes}}$ is exactly the Bayesian HMM $\mathbb{H}$ with a uniform prior on $\mathcal{A}_T$, where the experts Ξ have been identified with Bernoulli parameters $\mathcal{A}_T$. In Section 3.3 we will choose $\mathcal{A}_T$ differently from [Monteleoni and Jaakkola, 2003] based on our new discretisation scheme.

3.2.2 Regret Bounds

The following lemma will be our main tool to show regret bounds. It bounds the likelihood ratio between any two tracking HMMs in terms of the worst-case likelihood ratio of their corresponding Bernoulli HMMs. In other words, the lemma allows us to *lift* any uniform performance guarantees we may prove for Bernoulli HMMs to the level of tracking HMMs.

Lemma 3.1 (Lifting Lemma for Tracking)**.** *Suppose $\mathbb{B}_a$ and $\mathbb{B}_b$ are Bernoulli HMMs, and $\mathbb{B}_b(y^{T-1}) > 0$ for all binary sequences y^{T-1}. Then for any x^T*

$$\mathbb{S}_a(x^T) \leq \mathbb{S}_b(x^T) \max_{y^{T-1}} \frac{\mathbb{B}_a(y^{T-1})}{\mathbb{B}_b(y^{T-1})}.$$

By invoking this lemma with $\mathbb{S}_a = \mathbb{F}_{\hat{\alpha}(x^T)}$, where $\hat{\alpha}(x^T)$ is the best possible switching rate, we can obtain a bound on the regret for any tracking HMM with respect to the `Fixed-share` algorithm with optimally tuned parameter. This is the idea behind our main results, which appear as Theorem 3.1 below. The proof of the lemma uses the following more general lemma.

Lemma 3.2. *Let P and Q be distributions on countable space $\mathcal{Z} \times \Psi$ such that for all outcomes $\langle z, \psi \rangle$ we have $P(z \mid \psi) = Q(z \mid \psi)$ and $Q(\psi) > 0$. Then, for $z \in \mathcal{Z}$,*

$$P(z) \leq Q(z) \cdot \max_{\psi \in \Psi} \frac{P(\psi)}{Q(\psi)}.$$

Proof.

$$\begin{aligned} P(z) &= \sum_{\psi} P(\psi) P(z \mid \psi) \\ &\leq \max_{\psi} \frac{P(\psi)}{Q(\psi)} \sum_{\psi} Q(\psi) P(z \mid \psi) = Q(z) \max_{\psi} \frac{P(\psi)}{Q(\psi)}. \end{aligned}$$ □

Proof of Lemma 3.1. Let $Y_t = 1 - \mathbf{1}_{\{\xi_t\}}(\xi_{t+1})$ for $t = 1, \ldots, T$ indicate whether or not a switch occurs. Now let $\mathcal{Z} = \mathcal{X}^T$ and $\Psi = \{0,1\}^{T-1}$, and notice that for any $\left\langle x^T, y^{T-1} \right\rangle \in \mathcal{Z} \times \Psi$ we have

$$\begin{aligned} \mathbb{S}_a(x^T, y^{T-1}) &= \mathbb{F}(x^T \mid y^{T-1}) \mathbb{B}_a(y^{T-1}) \\ \mathbb{S}_b(x^T, y^{T-1}) &= \mathbb{F}(x^T \mid y^{T-1}) \mathbb{B}_b(y^{T-1}), \end{aligned}$$

where $\mathbb{F}(x^T \mid y^{T-1}) \equiv \mathbb{F}_\alpha(x^T \mid y^{T-1})$ denotes a conditional probability in the `Fixed-share` HMM that does not depend on α. Lemma 3.2 completes the proof. □

The lifting lemma is tight in the following sense. Consider two experts, whose predictions for all x^{t-1} are simply $P_1(X_t = 1 \mid x^{t-1}) = 1$ and $P_2(X_t = 0 \mid x^{t-1}) = 1$, respectively. Then any tracking HMM $\mathbb{S}$ with corresponding Bernoulli HMM $\mathbb{B}$ has $\mathbb{S}(x^T) = \mathbb{B}(y^{T-1})$, where $y_t = 1 - \mathbf{1}_{\{x_t\}}(x_{t+1})$ identifies whether the t-th and $(t+1)$-th outcomes are the same or not. Hence the regret is maximised for x^T such that the corresponding y^{T-1} maximises $\mathbb{B}_a(y^{T-1}) / \mathbb{B}_b(y^{T-1})$.

Section 3.3 introduces two new Bernoulli HMMs. We already mentioned the first one, $\mathbb{B}_{\text{Bayes}}$, in the example above. In Section 3.3.2 we

provide a uniform bound on its regret compared to any Bernoulli distribution. Then in Section 3.3.3 we define $\mathbb{B}_{\text{ro}}$, which does not require T to be known in advance, and extend the results from Section 3.3.2 to bound the regret of $\mathbb{B}_{\text{ro}}$. The `Refine-Online` algorithm is defined using this second Bernoulli HMM. Combining these results with Lemma 3.1 and the observation that $\mathbb{B}^{\alpha}_{\text{fixed}}(y^T) = P_\alpha(y^T)$ for all y^T, we directly obtain the main results of this chapter:

Theorem 3.1 (Learning the Switching Rate)**.** *Let $\mathbb{B}_{\text{Bayes}}$ be as in Definition 3.2 below. Then for any $\alpha \in [0,1]$ and any data x^T such that $T > 1$, the regret of $\mathbb{S}_{\text{Bayes}}$ compared to $\mathbb{F}_\alpha$ is bounded by*

$$\ln \frac{\mathbb{F}_\alpha(x^T)}{\mathbb{S}_{\text{Bayes}}(x^T)} \leq \frac{1}{2}\ln(T-1) + 2.8,$$

and the regret of $\mathbb{S}_{\text{ro}}$ is bounded by

$$\ln \frac{\mathbb{F}_\alpha(x^T)}{\mathbb{S}_{\text{ro}}(x^T)} \leq \log 3 \left(\frac{1}{2}\ln(T-1) + \ln\ln(T)\right) + 23.1.$$

(For $T = 1$, $\mathbb{S}_{\text{Bayes}}(x) = \mathbb{S}_{\text{ro}}(x) = \mathbb{F}_\alpha(x)$ for any x.)

While this theorem yields a bound for $\mathbb{S}_{\text{Bayes}}$ comparable to that given in [Monteleoni and Jaakkola, 2003], the analysis is different: in the end it is based on Lemma 3.1, which can only be usefully applied when good *uniform* bounds on the prior probability of the expert sequence, as established in Section 3.3.2, are available. In contrast, the analysis in [Monteleoni and Jaakkola, 2003] only requires a good bound on the Kullback-Leibler divergence $D(\hat{\alpha}\|\ddot{\alpha})$ between the *optimal* switching rate $\hat{\alpha}$ and the best discretised parameter $\ddot{\alpha} \in \mathcal{A}_T$. In other words, the only region where the discretisation precision actually matters is close to $\hat{\alpha}$. But their analysis does not readily generalise to other Bernoulli HMMs such as $\mathbb{B}_{\text{ro}}$.

3.3 Discretisation of Bernoulli Sources

In this section we define two Bernoulli HMMs, $\mathbb{B}_{\text{Bayes}}$ and $\mathbb{B}_{\text{ro}}$, and derive bounds on their worst-case regret. The first is based on a fixed discretisation of the set of Bernoulli distributions, where the optimal

number of discretisation levels depends on the total number of outcomes T, which therefore has to be known. The resulting tracking HMM, $\mathbb{S}_{\text{Bayes}}$, is similar to `Learn-`$\alpha$, but with the added advantage that the exact number and locations of the discretisation points are explicitly specified. Moreover, we obtain an explicit constant.

The analysis of $\mathbb{B}_{\text{Bayes}}$ is also an essential stepping stone to the specification of the second Bernoulli HMM $\mathbb{B}_{\text{ro}}$, whose discretisation of the set of Bernoulli distributions is not fixed; instead the discretisation is *refined* every time the number of outcomes gets large enough that it pays to do so.

Preliminaries As before, let P_α denote the Bernoulli distribution with $P_\alpha(Y = 1) = \alpha$. For any binary sequence y^T, the maximum likelihood parameter is $\hat{\alpha}(y^T) = T^{-1}\sum_{t=1}^{T} y_t$. When the data sequence is clear from context, we usually abbreviate $\hat{\alpha} \equiv \hat{\alpha}(y^T)$. The maximum likelihood is a sufficient statistic: for any α and T, the probability $P_\alpha(y^T)$ is completely determined by $\hat{\alpha}$. We therefore define $P_\alpha(\hat{\alpha}) := \alpha^{\hat{\alpha}}(1-\alpha)^{1-\hat{\alpha}}$, allowing any $\hat{\alpha} \in [0,1]$, not just rational values. Note that $T \ln P_\alpha(\hat{\alpha}) = \ln P_\alpha(y^T)$.

3.3.1 Discretisation

The analysis below is based on a different parametrisation of the Bernoulli distributions. For $\alpha \in [0,1]$ and $\phi \in [0, \pi/2]$, let $\phi(\alpha) = \arcsin\sqrt{\alpha}$ and $\alpha(\phi) = \sin^2\phi$. It is convenient to think of ϕ-parameters as points in the first quadrant of the unit circle. The parametrisation has many elegant properties; for example the Fisher information is constant. Similar *arcsine transformations* are well-known in the statistical literature [Anscombe, 1948, Freeman and Tukey, 1950]. In the following we will use $P_\alpha(\hat{\alpha})$ and $P_\phi(\hat{\phi})$ interchangeably, where the intended parametrisation should be clear from the parameter name and the context.

We now describe an explicit discretisation scheme for the ϕ-parameter of Bernoulli distributions that is especially easy to refine incrementally in online settings.

Definition 3.1 (k-Discretisation)**.** For $k \in \{1, 2, \ldots\}$ define the *k-discretisation* as the set $\mathcal{D}_k := \{\delta_k, 2\delta_k, 3\delta_k, \ldots, (2^k-1)\delta_k\} \cup \{\frac{1}{2}\delta_k, \pi/2 - \frac{1}{2}\delta_k\}$ of $2^k + 1$ discretisation points, where $\delta_k = \pi 2^{-k-1}$.

This is a uniform discretisation made slightly denser at the boundaries. The $(k+1)$-discretisation adds a new point midway between any two points in the k-discretisation, except at the boundaries, which require special care. Thus $\mathcal{D}_k \subset \mathcal{D}_{k+1}$, which will turn out to facilitate incremental refinement in the online setting.

Given k-discretisation $\mathcal{D}_k$, any point $\psi \in [0, \pi/2]$ has a set $N_k(\psi)$ of *neighbours* in $\mathcal{D}_k$, which is defined as

$$N_k(\psi) = \begin{cases} \{\phi_1\} & \text{if } \psi > \pi/2 - \delta_k/2, \\ \{\phi_2\} & \text{if } \psi < \delta_k/2, \\ \{\phi_1, \phi_2\} & \text{otherwise,} \end{cases}$$

where $\phi_1 = \max\{\phi \in \mathcal{D}_k \mid \phi \leq \psi\}$ and $\phi_2 = \min\{\phi \in \mathcal{D}_k \mid \phi \geq \psi\}$. (Note that $\phi_1 = \phi_2$ if $\psi \in \mathcal{D}_k$.)

3.3.2 The Offline Bernoulli HMM $\mathbb{B}_{\text{Bayes}}$

In the example above, we defined the offline Bernoulli HMM $\mathbb{B}_{\text{Bayes}}$ using an unspecified set $\mathcal{A}_T$ of discretisation points. We now complete the definition.

Definition 3.2. $\mathbb{B}_{\text{Bayes}}$ is the Bernoulli HMM as introduced in (3.2), defined with respect to $\mathcal{A}_T = \{\alpha(\phi) \mid \phi \in \mathcal{D}_{k(T)}\}$, where $k(T) = \left\lceil \frac{1}{2} \log(T\pi^2(2-\sqrt{2})) \right\rceil$.

As the number of transitions per time step equals $|\mathcal{A}_T|$ for this Bernoulli HMM, the forward algorithm for $\mathbb{B}_{\text{Bayes}}$ runs in $O(T\sqrt{T})$ time.

We proceed to analyse the regret of $\mathbb{B}_{\text{Bayes}}$ in the worst case over all possible binary sequences $y^T \in \{0,1\}^T$. The following lemma is at the basis for all of the following results. Its proof, and the proofs of the other results in this section, are deferred to Section 3.5.

Lemma 3.3 (Generalised Divergence Bound)**.** *Suppose ϕ_1, ϕ_2 and ϕ_3 all lie in $[0, \pi/4]$ and $\phi_2 > 0$. Then*

$$\begin{aligned} \ln \frac{P_{\phi_1}(\phi_3)}{P_{\phi_2}(\phi_3)} &= D(\phi_3 \| \phi_2) - D(\phi_3 \| \phi_1) \\ &\leq \begin{cases} 4(\phi_2 - \phi_1)(\phi_2 - \phi_3) & \text{if } \phi_3 \leq \phi_2, \\ 4(\phi_2 - \phi_1)(\phi_2 - \phi_3)\dfrac{\phi_3}{\phi_2} & \text{otherwise.} \end{cases} \end{aligned} \tag{3.3}$$

Note that by symmetry in $\pi/4$ the lemma can also be applied to $\phi_i' = \pi/2 - \phi_i$ for $i = 1, 2, 3$. Although it provides a bound on the Kullback-Leibler divergence, which is an expected quantity, we use it to prove results on individual sequence regret. In particular, Lemma 3.3 will typically be applied with ϕ_3 set to the maximum likelihood $\hat{\phi}$ for some binary sequence. As a notational reminder, ϕ_3 will be called $\hat{\phi}$ in the remainder.

The following consequence of Lemma 3.3 is an important intermediate result. It expresses that the regret per outcome of using the best discretisation point rather than the maximum likelihood is $O(\delta_k^2)$, which means that $O(\sqrt{T})$ uniformly spaced discretisation points suffice to achieve an $O(1)$ overall worst-case regret. Using the ϕ-parametrisation is crucial; in the α-parametrisation the discretisation points must be packed extra densely near the boundaries of the parameter space.

Lemma 3.4 (Discretisation Lemma)**.** *For any $\hat{\phi} \in [0, \pi/2]$ and $\phi \in (0, \pi/2)$ it holds that*

$$\min_{\phi \in N_k(\hat{\phi})} \ln \frac{P_{\hat{\phi}}(\hat{\phi})}{P_{\phi}(\hat{\phi})} \leq (8 - 4\sqrt{2})\delta_k^2 \leq 2.4\ \delta_k^2.$$

Specifically, for $\mathbb{B}_{\text{Bayes}}$ we obtain the following worst-case regret bound.

Theorem 3.2 (Offline Discretisation)**.** *For any binary sequence $y^T \in \{0,1\}^T$ and any $\alpha \in [0, 1]$*

$$\ln \frac{P_\alpha(y^T)}{\mathbb{B}_{\text{Bayes}}(y^T)} \leq \frac{1}{2} \ln T + 2.8.$$

3.3.3 The Online Bernoulli HMM $\mathbb{B}_{\text{ro}}$

We shall now define the remaining properties of the Refine-Online Bernoulli HMM, $\mathbb{B}_{\text{ro}}$, using $\mathcal{D}_k$ as before. But since we do not know T, rather than choosing a fixed k as a function of T, we let k increase by one every time the precision threatens to become insufficient, roughly doubling the number of discretisation points. The critical step in the definition of $\mathbb{B}_{\text{ro}}$ will describe how to patch things up whenever k increases.

Our approach is more subtle than the *doubling trick*, which is often used to deal with unknown T [Cesa-Bianchi and Lugosi, 2006]. Naive doubling can be done in two ways. The simplest is to restart

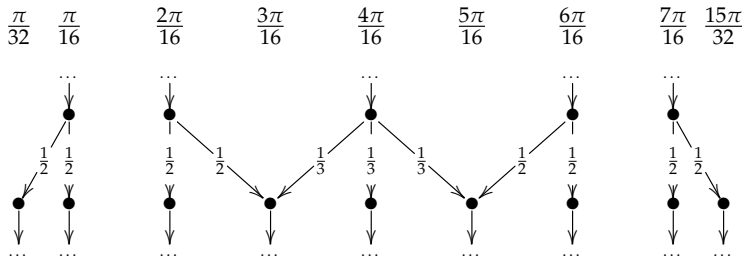


Figure 3.2: Refinement from $\mathcal{D}_2$ to $\mathcal{D}_3$.

the algorithm completely each time the precision needs to be increased. But then the Bernoulli parameter has to be relearned in each segment, which results in a significantly worse loss bound of order $O((\ln T)^2)$. Alternatively, one might revisit previous data and continue by setting the algorithm's weights as if the increased precision had been used from the start. But this requires the algorithm to store all data indefinitely; moreover, we have not been able to improve our loss bound using this approach. In the following we therefore suggest a more advanced way of doubling, which redistributes the weights of the algorithm without looking at old data whenever the precision is increased.

We first define $\mathbb{B}^k_{\mathrm{ro}}$ with respect to a function $k\colon \mathbb{Z}^+ \to \mathbb{Z}^+$, called the *discretisation function*. It identifies the discretisation set $\mathcal{D}_{k(t)}$ to be used at time t, and should have the property that $k(t+1) = k(t)$ or $k(t+1) = k(t)+1$ for all t. Thus $\mathcal{A}_t = \{\alpha(\phi) \mid \phi \in \mathcal{D}_{k(t)}\}$. The discretisation function for `Refine-Online` is

$$\kappa(t) = \left\lfloor \frac{1}{2}\log t + \log\log(t+1) \right\rfloor + 1,$$

and we simply write $\mathbb{B}_{\mathrm{ro}}$ for $\mathbb{B}^{\kappa}_{\mathrm{ro}}$.

The initial weights of the states are $\mathbb{B}^k_{\mathrm{ro}}(\langle \alpha_1, 1\rangle) = 1/|\mathcal{D}_{k(1)}|$. It remains to define the transition probabilities between states. For consecutive times t and $t+1$ when the discretisation does not change, i.e. $k(t) = k(t+1)$, these transitions are similar to those for the Bayesian Bernoulli HMM in (3.2); for times when the discretisation does change, the probabilities are given by a *refinement function* $d_k\colon \mathcal{D}_k \times \mathcal{D}_{k+1} \to$

$[0,1]$. Thus,

$$\mathbb{B}^k_{\mathrm{ro}}(\langle \alpha_{t+1}, t+1\rangle \mid \langle \alpha_t, t\rangle) = \begin{cases} \mathbf{1}_{\{\alpha_t\}}(\alpha_{t+1}) & \text{if } k(t) = k(t+1), \\ d_{k(t)}(\phi(\alpha_t), \phi(\alpha_{t+1})) & \text{otherwise.} \end{cases}$$

The refinement function d_k, which determines our patch-up strategy, is chosen such that ϕ_{t+1} gets some mass from each of its neighbours in $N_{k(t)}(\phi_{t+1})$:

$$d_k(\phi_t, \phi_{t+1}) = \mathbf{1}_{N_k(\phi_{t+1})}(\phi_t) \cdot \begin{cases} \frac{1}{2} & \text{if } \phi_t \le \delta_k \text{ or } \phi_t \ge \frac{1}{2}\pi - \delta_k, \\ \frac{1}{3} & \text{otherwise.} \end{cases}$$

The refinement function is illustrated by Figure 3.2 for $k = 2$, but note that as $k(t)$ gets larger, the case that $d_k(\phi_t, \phi_{t+1}) = 1/3$ becomes most important. Also note there are at most three transitions for each discretisation point per time step. The forward algorithm therefore runs in time proportional to $\sum_{t=1}^{T} |\mathcal{D}_{k(t)}| \le T\,|\mathcal{D}_{k(T)}|$. In particular for $\mathbb{B}_{\mathrm{ro}}$ ($k = \kappa$) its running time is $O(T\sqrt{T}\log T)$.

While it may seem redundant to allow for converging paths in the HMM, we do need such a structure for the proof of the lemma below, which bounds the weights of the newly introduced discretisation points. The idea is to compare the weight that is accumulated in any state $\langle \alpha_t, t\rangle$ after observing y^t, to $P_{\alpha_t}(y^t)$. Let $t(k) = \min\{t \in \mathbb{Z}^+ \mid k(t) = k\}$ be the first time at which the k-discretisation is used. If the discretisation function k were strictly increasing, this would be its inverse.

Lemma 3.5 (Refinement Lemma). *For any $y^t \in \{0,1\}^t$, any $\phi \in \mathcal{D}_{k(t)}$ it holds that*

$$\ln \frac{P_\phi(y^t)}{\mathbb{B}^k_{\mathrm{ro}}(y^t, \langle \alpha(\phi), t\rangle)} \le \ln |\mathcal{D}_{k(1)}| + \sum_{k=k(1)+1}^{k(t)} \ln 3 + (4 - 2\sqrt{2})\pi^2 \frac{t(k) - 1}{4^k}. \tag{3.4}$$

In particular for the discretisation function κ we get

$$\ln \frac{P_\phi(y^t)}{\mathbb{B}_{\mathrm{ro}}(y^t, \langle \alpha(\phi), t\rangle)} \le \log 3 \left(\frac{1}{2} \ln t + \ln \ln(t+1) \right) + 20.7.$$

Using this lemma it is not hard to provide a worst-case regret bound for $\mathbb{B}_{\mathrm{ro}}$.

Theorem 3.3 (Online Discretisation). *For any binary sequence $y^t \in \{0,1\}^t$ and any $\alpha \in [0,1]$*

$$\ln \frac{P_\alpha(y^t)}{\mathbb{B}_{\mathrm{ro}}(y^t)} \leq \log 3 \left(\frac{1}{2} \ln t + \ln \ln(t+1) \right) + 23.1.$$

Here the constant is the sum of the constants appearing in Lemmas 3.4 and 3.5. The proof of this theorem is based on the regret of the discretisation point $\ddot{\phi}(y^t) \in \mathcal{D}_{k(t)}$ that is closest to the unconstrained maximum likelihood $\hat{\phi}(y^t)$. There are $O(\log t)$ discretisation points sufficiently close to $\hat{\phi}(y^t)$. Taking this into account would result in an improved constant in front of the $\ln \ln(t+1)$ term, but the term would not vanish and the proof would become more complex.

3.4 Conclusion

We have presented a new discretisation scheme for Bernoulli sources that achieves a regret bound of $\frac{1}{2} \ln T + 2.8$ nats if the final number of outcomes, T, is known in advance, but unlike the approach in [Monteleoni and Jaakkola, 2003] specifies the exact number and positions of the discretisation points explicitly. This scheme is most useful, however, when T is not known in advance: in Section 3.3.3 the HMM $\mathbb{B}_{\mathrm{ro}}$ was presented that achieves a regret of $\frac{1}{2} \log 3 \ln T + \log 3 \ln \ln(T+1) + 23.1$ nats without knowing T in advance. The predictions of $\mathbb{B}_{\mathrm{ro}}$ can be computed in $O(T\sqrt{T} \log T)$ time using the standard forward algorithm for HMMs.

Our interest in Bernoulli sources stems from Lemma 3.1, which shows that these bounds directly translate into regret bounds for learning the switching rate for the `Fixed-share` algorithm. As discussed in Section 3.2.1, running times also carry over. We call the new algorithm for the case where T is not known `Refine-Online`.

Analogues to Lemma 3.1 may easily be proved for any expert algorithm that involves a repeated binary choice with fixed probability, like *elementwise mixtures* [Koolen and de Rooij, 2008a].

Future Research The worst-case regret for Bernoulli sources is $\frac{1}{2}\log T + O(1)$ [Cesa-Bianchi and Lugosi, 2006, Thm 9.2]. This provides a lower bound on the worst-case regret for tracking HMMs, because Lemma 3.1 is tight. The lower bound is achieved by $\mathbb{S}_{\text{Bayes}}$, but for $\mathbb{S}_{\text{ro}}$ a $\log 3$ factor appears. This factor can be explained as follows. When the discretisation is refined, each new point gets mass from two neighbours, but our analysis in Lemma 3.4 only takes the best neighbour into account. It is an interesting open question whether the optimal bound could be achieved, at least up to $O(\log\log T)$, by improving either the refinement function or the analysis.

3.5 Proofs

Generalised Divergence Bound (Lemma 3.3) The equality follows by rewriting definitions. The inequality is proved as follows. For any concave function f with derivative f', and any x and y, it holds that

$$(x-y)f'(x) \leq f(x) - f(y) \leq (x-y)f'(y). \tag{3.5}$$

In particular for $\ln P_\phi(\phi_3)$ as a function of ϕ:

$$\begin{aligned}\ln \frac{P_{\phi_1}(\phi_3)}{P_{\phi_2}(\phi_3)} &\leq (\phi_1-\phi_2)\left(2\alpha_3\frac{\cos\phi_2}{\sin\phi_2} - 2(1-\alpha_3)\frac{\sin\phi_2}{\cos\phi_2}\right)\\ &= 2(\phi_1-\phi_2)\frac{\cos^2\phi_2 - \cos^2\phi_3}{\sin\phi_2\cos\phi_2}.\end{aligned} \tag{3.6}$$

Since $\cos^2\phi$ is a concave function of ϕ as well, we can use (3.5) once more to find

$$\begin{aligned}-2(\phi_2-\phi_3)\sin\phi_2\cos\phi_2 \leq \cos^2\phi_2 - \cos^2\phi_3 \\ \leq -2(\phi_2-\phi_3)\sin\phi_3\cos\phi_3.\end{aligned} \tag{3.7}$$

If $\phi_2 - \phi_3 \geq 0$, then plugging the left-hand side into (3.6) gives the first case of (3.3). For $\phi_2 - \phi_3 < 0$ we first combine the inequality on the right hand side of (3.7) with (3.6) to find

$$\ln \frac{P_{\phi_1}(\phi_3)}{P_{\phi_2}(\phi_3)} \leq 4(\phi_1-\phi_2)(\phi_3-\phi_2)\frac{\sin\phi_3\cos\phi_3}{\sin\phi_2\cos\phi_2}. \tag{3.8}$$

As $\sin x \cos x = \sin 2x$ and $\sin x$ is concave on $[0, \pi/2]$, we also get by (3.5) that

$$\frac{\sin \phi_3 \cos \phi_3}{\sin \phi_2 \cos \phi_2} \leq 1 + \frac{2(\phi_3 - \phi_2)}{\tan(2\phi_2)} \leq \frac{\phi_3}{\phi_2},$$

where the second inequality follows by $\tan x \geq x$ for $x \in [0, \pi/2]$. With (3.8) this completes the proof. □

Discretisation Lemma (Lemma 3.4) We first show that for any $0 < \phi_1 \leq \hat{\phi} \leq \phi_2 \leq \pi/4$ it holds that

$$\min_{\phi \in \{\phi_1, \phi_2\}} \ln \frac{P_{\hat{\phi}}(\hat{\phi})}{P_{\phi}(\hat{\phi})} \leq 4(\phi_2 - \sqrt{\phi_1 \phi_2})^2. \tag{3.9}$$

This follows by relaxing Lemma 3.3 to get

$$\ln \frac{P_{\hat{\phi}}(\hat{\phi})}{P_{\phi_1}(\hat{\phi})} \leq 4(\phi_1 - \hat{\phi})^2 (\hat{\phi}/\phi_1)^2,$$

which is strictly increasing in $\hat{\phi}$, and

$$\ln \frac{P_{\hat{\phi}}(\hat{\phi})}{P_{\phi_2}(\hat{\phi})} \leq 4(\phi_2 - \hat{\phi})^2,$$

which is strictly decreasing. At the maximising $\hat{\phi} = \sqrt{\phi_1 \phi_2}$ the bounds are equal. Substitution completes the proof of (3.9).

To prove Lemma 3.4, assume without loss of generality that $\hat{\phi} \leq \pi/4$; the other case is symmetric. Then $\phi \leq \pi/4$ for all $\phi \in N_k(\hat{\phi})$. If $N_k(\hat{\phi}) = \{\hat{\phi}\}$, the lemma is trivially true. If $N_k(\hat{\phi}) = \{\delta_k/2\}$, then $\hat{\phi} \leq \delta_k/2$ and from Lemma 3.3 we get

$$\ln \frac{P_{\hat{\phi}}(\hat{\phi})}{P_{\delta_k/2}(\hat{\phi})} \leq 4(\frac{1}{2}\delta_k - \hat{\phi})^2 \leq \delta_k^2.$$

If $N_k(\hat{\phi}) = \{\frac{1}{2}\delta_k, \delta_k\}$ we similarly obtain a bound of δ_k^2. Finally, suppose that $N_k(\hat{\phi}) = \{i\delta_k, (i+1)\delta_k\}$ for some integer $i \geq 1$. Then application of (3.9) yields

$$\min_{\phi \in \{i\delta_k, (i+1)\delta_k\}} \ln \frac{P_{\hat{\phi}}(\hat{\phi})}{P_{\phi}(\hat{\phi})} \leq 4\Big((i+1) - \sqrt{i(i+1)}\Big)\delta_k^2,$$

which is maximised by $i = 1$. □

Offline Discretisation (Theorem 3.2) Let $\hat{\phi}$ denote the maximum likelihood and $\ddot{\phi} = \arg\max_{\phi \in \mathcal{D}_k} P_\phi(y^T)$ denote the maximum likelihood in $\mathcal{D}_k$. The theorem follows by $-\ln \mathbb{B}_{\text{Bayes}}(y^T) \leq -\ln P_{\ddot{\phi}}(y^T) - \ln w(\ddot{\phi})$ and Lemma 3.4. □

Lemma 3.6. *Suppose that* $0 < \phi_1 \leq \phi_2 \leq \pi/4$ *and define* $\psi = \frac{1}{2}(\phi_1 + \phi_2)$. *Then for any* $\hat{\phi} \in [0, \pi/2]$,

$$\min_{\phi \in \{\phi_1, \phi_2\}} \ln \frac{P_\psi(\hat{\phi})}{P_\phi(\hat{\phi})} \leq 2(\phi_2 - \phi_1)(\phi_2 - \sqrt{\phi_1 \phi_2}). \tag{3.10}$$

Proof. As $\ln P_\phi(\hat{\phi})$ is a concave function of ϕ achieving its maximum at $\phi = \hat{\phi}$, we have for $\hat{\phi} < \phi_1$ or $\hat{\phi} > \phi_2$ that $\min_{\phi \in \{\phi_1, \phi_2\}} \ln P_\psi(\hat{\phi}) / P_\phi(\hat{\phi}) \leq 0$, such that (3.10) is satisfied. Therefore assume without loss of generality that $\phi_1 \leq \hat{\phi} \leq \phi_2$. At the worst-case $\hat{\phi}$, the bounds from Lemma 3.3 must be equal; solving yields $\hat{\phi} = \sqrt{\phi_1 \phi_2}$. Substitution in one of the bounds completes the proof. □

Lemma 3.7. *For all* $\psi \in \mathcal{D}_{k+1}$ *and any* $\hat{\phi} \in [0, \pi/2]$,

$$\min_{\phi \in N_k(\psi)} \ln \frac{P_\psi(\hat{\phi})}{P_\phi(\hat{\phi})} \leq (4 - 2\sqrt{2})\delta_k^2 = \frac{(4 - 2\sqrt{2})\pi^2}{4^{k+1}}. \tag{3.11}$$

Proof. Assume without loss of generality that $\psi < \pi/4$. Then $\phi \leq \pi/4$ for all $\phi \in N_k(\psi)$. If $\psi \in \mathcal{D}_k$, then the lemma is trivially true. If $\psi = \delta_{k+1}/2$, then $N_k(\psi) = \{\delta_k/2\}$, and as $\ln P_\phi(\hat{\phi})$ is concave in ϕ and achieves its maximum at $\phi = \hat{\phi}$, (3.11) is satisfied if $\hat{\phi} > \delta_k/2$. If $\hat{\phi} \leq \delta_k/2$ it follows by Lemma 3.3 that

$$\ln \frac{P_{\delta_{k+1}/2}(\hat{\phi})}{P_{\delta_k/2}(\hat{\phi})} \leq 4(\delta_{k+1}/2)(\delta_k/2 - \hat{\phi}(y^t)) \leq \frac{1}{2}\delta_k^2.$$

If neither of these cases apply, we must have $N_k(\psi) = \{\phi_1, \phi_2\}$ with $\phi_1 = i\delta_k$ and $\phi_2 = (i+1)\delta_k$ for some integer $i \geq 1$, and $\psi = (\phi_1 + \phi_2)/2$. In that case we apply Lemma 3.6 to find

$$\min_{\phi \in N_k(\psi)} \ln \frac{P_\psi(\hat{\phi})}{P_\phi(\hat{\phi})} \leq 2\delta_k^2(i + 1 - \sqrt{i(i+1)}),$$

which is maximised by $i = 1$. □

Refinement Lemma (Lemma 3.5) Abbreviate $\langle \alpha(\phi), t \rangle$ to $\langle \phi, t \rangle$ and let $b(t)$ denote the right-hand side of (3.4). The proof of the first part of the lemma is by induction on t. The case $t = 1$, for which $b(t) = \ln |\mathcal{D}_{k(1)}|$, is verified by noting that $\mathbb{B}_{\mathrm{ro}}(y^1, \langle \phi, 1 \rangle) = P_\phi(y^1)/|\mathcal{D}_{k(1)}|$. Suppose the bound is valid for some t. To show that it is also valid for $t+1$, using that

$$\begin{aligned}
&\ln \frac{P_{\phi_{t+1}}(y^{t+1})}{\mathbb{B}_{\mathrm{ro}}(y^{t+1}, \langle \phi_{t+1}, t+1 \rangle)} - \ln \frac{P_{\phi_{t+1}}(y^t)}{\mathbb{B}_{\mathrm{ro}}(y^t, \langle \phi_t, t \rangle)} \\
&\quad \leq \min_{\phi_t \in \mathcal{D}_{k(t)}} \ln \frac{P_{\phi_{t+1}}(y_{t+1})}{\mathbb{B}_{\mathrm{ro}}(y_{t+1}, \langle \phi_{t+1}, t+1 \rangle \mid y^t, \langle \phi_t, t \rangle)} \\
&\quad = \min_{\phi_t \in \mathcal{D}_{k(t)}} -\ln \mathbb{B}_{\mathrm{ro}}(\langle \phi_{t+1}, t+1 \rangle \mid \langle \phi_t, t \rangle).
\end{aligned}$$

In case $k(t+1) = k(t)$ the bound does not change (i.e. $b(t+1) = b(t)$), because for $\phi_t = \phi_{t+1} \in \mathcal{D}_{k(t)}$ it holds that $\mathbb{B}_{\mathrm{ro}}(\langle \phi_{t+1}, t+1 \rangle \mid \langle \phi_t, t \rangle) = 1$, and by induction $\ln P_{\phi_t}(y^t) - \ln \mathbb{B}_{\mathrm{ro}}(y^t, \langle \phi_t, t \rangle) \leq b(t)$. Now suppose that $k(t+1) = k(t) + 1$. Then

$$\begin{aligned}
&\min_{\phi_t \in \mathcal{D}_{k(t)}} -\ln \mathbb{B}_{\mathrm{ro}}(\langle \phi_{t+1}, t+1 \rangle \mid \langle \phi_t, t \rangle) + \ln \frac{P_{\phi_{t+1}}(y^t)}{\mathbb{B}_{\mathrm{ro}}(y^t, \langle \phi_t, t \rangle)} \\
&\quad = \min_{\phi_t \in N_{k(t)}(\phi_{t+1})} -\ln d_{k(t)}(\phi_t, \phi_{t+1}) + \ln \frac{P_{\phi_{t+1}}(y^t)}{\mathbb{B}_{\mathrm{ro}}(y^t, \langle \phi_t, t \rangle)} \\
&\quad \leq \ln 3 + \min_{\phi_t \in N_{k(t)}(\phi_{t+1})} \ln \frac{P_{\phi_{t+1}}(y^t)}{P_{\phi_t}(y^t)} + b(t) \\
&\quad \leq \ln 3 + (4 - 2\sqrt{2})\pi^2 \frac{t}{4^{k(t)+1}} + b(t) = b(t+1),
\end{aligned}$$

where the first inequality holds by induction and the last inequality follows from Lemma 3.7.

For the second part of the lemma we bound $t(k)$ using

$$\sqrt{t(k)} \log(t(k)+1) \leq 2^k \leq 2\sqrt{t(k)} \log(t(k)+1). \tag{3.12}$$

From the left-hand side of (3.12) we get

$$\sqrt{t(k)} \leq \frac{2^k}{\log(t(k)+1)} \leq \frac{2^k}{\frac{1}{2}\log t(k) + \log\log(t(k)+1)}.$$

(We omit the tedious proof of the last inequality.) Together with the right-hand side of (3.12) it follows that $\sqrt{t(k)} \leq 2^k(k-1)^{-1}$, which implies $t(k) \leq 4^k(k-1)^{-2} \leq (4^k+1)(k-1)^{-2}$. The result follows by plugging this bound into (3.4). □

Online Discretisation (Theorem 3.3) Fix an arbitrary sequence y^t, and define the global maximum likelihood $\hat{\phi} = \hat{\phi}(y^t)$ and the nearest discretisation point $\ddot{\phi} = \arg\max_{\phi \in \mathcal{D}_{k(t)}} P_\phi(y^t)$. Then

$$\ln \frac{P_{\hat{\phi}}(y^t)}{\mathbb{B}_{\mathrm{ro}}(y^t)} \leq t \ln \frac{P_{\hat{\phi}}(\hat{\phi})}{P_{\ddot{\phi}}(\hat{\phi})} + \ln \frac{P_{\ddot{\phi}}(y^t)}{\mathbb{B}_{\mathrm{ro}}(y^t, \langle \alpha(\ddot{\phi}), t \rangle)}.$$

The latter two terms can be bounded using Lemmas 3.4 and 3.5, respectively. □

Chapter 4

Switching between Hidden Markov Models using Fixed-share

In prediction with expert advice the goal is to design online prediction algorithms that achieve small regret (additional loss on the whole data) compared to a reference scheme. In the simplest such scheme one compares to the loss of the best expert in hindsight. A more ambitious goal is to split the data into segments and compare to the best expert on each segment. This is appropriate if the nature of the data changes between segments. The standard `Fixed-share` algorithm is fast and achieves small regret compared to this scheme.

`Fixed-share` treats the experts as black boxes: there are no assumptions about how they generate their predictions. But if the experts are learning, the following question arises: should the experts learn from all data or only from data in their own segment? The original algorithm naturally addresses the first case. Here we consider the second option, which is more appropriate exactly when the nature of the data changes between segments. In general extending `Fixed-share` to this second case will slow it down by a factor of T on T outcomes. We show, however, that no such slowdown is necessary if the experts are hidden Markov models.

4.1 Introduction

In *prediction with expert advice* [Cesa-Bianchi and Lugosi, 2006] a sequence of outcomes $x_1, x_2, \ldots$ needs to be predicted, one outcome at a time. Thus, prediction proceeds in rounds: in each round we first consult a set of experts, who give us their predictions. (We use the word *expert* for any source of predictions that is available to us as in-

put.) Then we make our own prediction and incur some loss based on the discrepancy between our prediction and the actual outcome. Predictions may for example be in the form of a probability distribution on outcomes. Loss may be logarithmic loss, i.e. the negative logarithm of the probability assigned to the outcome that actually occurs. The goal is to minimise our *regret*, which is the difference between our own cumulative loss on the whole data and the cumulative loss of a *reference scheme*, which typically involves tuned parameter settings unknown to us when we make our predictions. For the reference scheme there are several options; we may, for example, compare ourselves to the cumulative loss of the best expert in hindsight (after observing the data). A more ambitious scheme, called *tracking the best expert*, is addressed by the `Fixed-share` algorithm of Herbster and Warmuth [1998].

4.1.1 Tracking the Best Expert

In tracking the best expert (TBE), the goal is to achieve small regret compared to the following reference scheme:

(a) Split the data into segments.

(b) Select an expert for each segment.

(c) Sum the loss of the selected experts on their segments.

This reference scheme is appropriate if the nature of the data changes between segments. It is harder than comparing to the single best expert in hindsight, because now there are more unknowns: both the segmentation (step a) and the reference experts (step b) are unknown when we make our predictions. In particular the reference experts may be the best experts in hindsight for their assigned segments.

The `Fixed-share` algorithm is efficient and achieves small regret (see Theorem 4.1 below) compared to the TBE reference scheme. Given the predictions of the experts, the algorithm's running time is linear in the number of outcomes and linear in the number of experts. Problem solved. Or is it?

4.1.2 Learning Experts

In this chapter we take another look at the TBE reference scheme for *learning experts* and ask: if an expert is selected for some segment, then

should the expert learn from all data or only from the data in that segment?

We may assume that the experts do not know the segmentation chosen in step a of the reference scheme. (Otherwise, why not just ask them?) Hence if we treat the experts as black boxes and only ask for their prediction at each time step as in [Herbster and Warmuth, 1998], it is natural that they learn from all data. We call this the *standard* interpretation of the TBE reference scheme (S-TBE).

However, as the following example will illustrate, it may be beneficial if experts learn only from the segment for which they are selected, because they may get confused by data in other segments that follow a different pattern. We call this the *local learners* interpretation of tracking the best expert (LL-TBE). As a slight complication, it will turn out that in LL-TBE we have a further choice: whether to tell a learning expert the timing of its segment or not, which generally makes a difference. When segment timing is preserved, we call the resulting reference scheme *sleeping LL-TBE*; when segment timing is *not* preserved we call the reference scheme *freezing LL-TBE*. The next example illustrates that S-TBE and the two variants of LL-TBE are very different reference schemes indeed.

Example: Drifting Mean In applications one would usually build up complicated prediction strategies from simpler ones in a hierarchical fashion. For example, let us first define simple static experts, parametrised by $\mu \in \mathbb{R}$, which predict according to a standard normal distribution with mean μ in each round. Now define a learning expert $\mathsf{DM}[\theta]$ that has a stochastic model for the (unobservable) drift of μ over time. This *drifting mean* learning expert predicts according to a hidden Markov model in which the hidden state at time t is μ_t and the production probability of an outcome given μ_t is determined by the simple expert with parameter μ_t. Initially, $\mu_1 = 0$ with probability one. Then $\mu_{t+1} = \mu_t + 1$ with probability θ and $\mu_{t+1} = \mu_t$ with probability $1 - \theta$ for some fixed parameter θ. (See Figure 4.1.)

The expert $\mathsf{DM}[\theta]$ may be said to be learning, because its posterior distribution of μ_t given outcomes $x_1, \ldots, x_{t-1}$ indicates how much credibility the expert assigns to each value of μ_t: high weight on, say, $\mu_t = 3$ indicates that $\mathsf{DM}[\theta]$ considers it likely for $\mu_t = 3$ to give the best prediction for x_t.

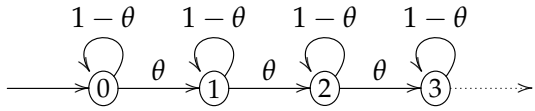
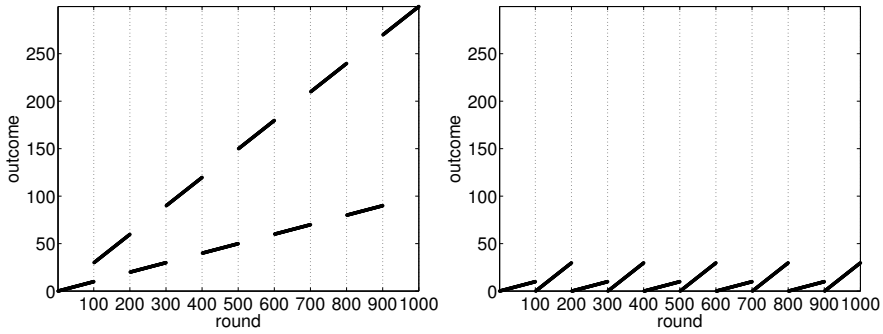
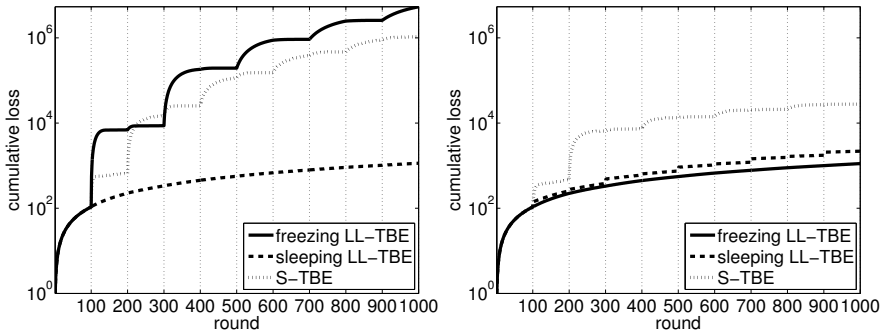


Figure 4.1: State transitions for learning expert DM[θ], which learns a drifting mean



(a) Suitable sleeping LL-TBE data

(b) Suitable freezing LL-TBE data



(c) Cumul. loss on data (a)

(d) Cumul. loss on data (b)

Figure 4.2: The difference between S-TBE and the two LL-TBE reference schemes. Note the logarithmic scale of the y-axis in (c) and (d)!

Figures 4.2a and 4.2b plot two artificial data sets. For Figure 4.2a sleeping LL-TBE is appropriate, for Figure 4.2b freezing LL-TBE is more suitable. The data consist of 10 segments of 100 outcomes. In each segment the outcomes are increasing deterministically at a rate of either 0.1 or 0.3 per outcome. Note that for the freezing data all segments start from 0, whereas for sleeping any segment looks like the proces that generated it started at 0 at time 1, but went unobserved for a while.

Figures 4.2c and 4.2d show the cumulative log(arithmic) loss for all three TBE reference schemes. Note that the difference between the schemes is so large that their losses had to be plotted on a logarithmic scale. In each case we consider two experts: $\mathsf{DM}[0.1]$ and $\mathsf{DM}[0.3]$ and use the expert $\mathsf{DM}[\theta]$ for any segment with rate θ. The difference between the three schemes lies in which data is used by $\mathsf{DM}[\theta]$ to learn from. In the S-TBE scheme $\mathsf{DM}[\theta]$ is shown all the data, even those outside the segment it has to predict. In the two LL-TBE schemes, on the other hand, a fresh copy of $\mathsf{DM}[\theta]$ only sees the data in the segment for which it is selected: for freezing LL-TBE, $\mathsf{DM}[\theta]$ predicts as if the current segment is the only data; for sleeping LL-TBE, $\mathsf{DM}[\theta]$ knows the timing of the segment it is predicting, and treats all samples preceding that segment as unobserved. Thus in sleeping LL-TBE the original timing of the segments is preserved, while in freezing LL-TBE it is lost.

We see that for the sleeping data the sleeping LL-TBE reference scheme has much smaller loss than the other two schemes. And for the freezing data the freezing LL-TBE scheme has the smallest loss by far. (Mind the logarithmic scale of the y-axis, which puts the loss of sleeping LL-TBE deceptively close to the loss of freezing LL-TBE in Figure 4.2d: a constant offset indicates a fixed multiplicative overhead.) In both cases the reason for the large differences between the reference schemes is that $\mathsf{DM}[\theta]$ gets confused if it learns from the wrong data.

4.1.3 Expert Hidden Markov Models

The learning expert $\mathsf{DM}[\theta]$ in the example above is a hidden Markov model in which the production probabilities (of outcomes given the state) depend on lower-level base experts. In general such prediction strategies are called *expert hidden Markov models* (EHMMs). The use of EHMMs is not restricted to describing learning experts. For example, many algorithms for prediction with expert advice, including FS itself, can be represented as EHMMs (see Koolen and De Rooij [2008a] and

its references, and Monteleoni and Jaakkola [2003]). In addition any ordinary HMM is trivially an EHMM: just introduce lower-level base experts for its production probabilities. Not every algorithm can be represented as an EHMM, however. The `Follow-the-perturbed-leader` algorithm by Hannan [1957] and `Variable-share` by Herbster and Warmuth [1998], for instance, are exceptions.

4.1.4 Fixed-share for Learning Experts

LL-TBE Requires More Information The example above shows that there is a large difference between S-TBE and the sleeping or freezing LL-TBE reference schemes. One may therefore wonder whether there exists an algorithm that achieves small regret compared to LL-TBE. Unfortunately, no algorithm will be able to do the job without additional knowledge about the learning experts. To see this, note that the reference scheme may split the data into segments in any way it sees fit. But black-box experts are not telling us what their predictions would be for any possible segmentation; they only give us a single prediction each round. Therefore, even if we knew the segmentation and the selected expert for each segment, we still would have insufficient information to achieve the reference scheme. The only way to address this problem is to get more information about the learning experts. This information should have an efficient representation and should somehow tell us what the learning experts would predict for any possible segmentation.

Copying Experts is Less Efficient The straight-forward approach would be to introduce a fresh copy of each expert for each possible start of a new segment and run the original `Fixed-share` algorithm on the resulting enriched set of experts. But then the number of experts would grow linearly with the number of rounds, and consequently the total running time would go up from linear to quadratic in the number of outcomes. As this makes the difference between an online algorithm that can run forever and an algorithm that effectively comes to a stop after, say, 10^5 outcomes, it is worth seeing whether such an increase in running time is really unavoidable.

EHMMs: the Efficient Special Case As we will show, it turns out there is a special class of learning experts for which no increase in

running time is necessary. These are the learning experts that can be described in EHMM form. Although this excludes learning experts that for example implement `Follow-the-perturbed-leader`, the class of EHMMs is still rich enough to be of interest, if only because it includes all ordinary HMMs. In the interpretation of the two LL-TBE reference schemes for learning experts in EHMM form, we do need to be careful if the base experts in the EHMMs are learning themselves: because we make no assumptions about the base experts, they always learn from all the data.

Main Result: Achieving LL-TBE Efficiently We present two new algorithms: $\mathsf{FS}^{\mathsf{sl}}$ for sleeping LL-TBE and $\mathsf{FS}^{\mathsf{fr}}$ for freezing LL-TBE, which both generalise FS. We show that these algorithms achieve the same regret bound compared to their respective LL-TBE reference schemes as FS achieves compared to the S-TBE reference scheme. In addition, $\mathsf{FS}^{\mathsf{sl}}$ runs equally fast as the original `Fixed-share` algorithm; for $\mathsf{FS}^{\mathsf{fr}}$ no slowdown occurs either if the EHMMs for the learning experts have a finite number of hidden states, otherwise it is typically still faster than just copying the experts.

Like `Fixed-share`, our new algorithms can be represented as EHMMs. In fact, we will build up both algorithms by describing how to combine the EHMMs for the learning experts, which the algorithms get as inputs, into a single larger EHMM. Apart from introducing the LL-TBE reference scheme, this construction is our main result: regret bounds follow from the EHMM representations using methods described in [Koolen and De Rooij, 2008a], and the algorithms are simply instances of the forward algorithm for EHMMs.

4.1.5 Overview

In the next section we introduce the notation for prediction with expert advice that is used in this chapter. Then Section 4.3 reviews EHMMs, including the representation of FS as an EHMM. It is shown how the standard regret bound for FS by Herbster and Warmuth [1998] can be proved using this representation. In Section 4.4 we formally define the freezing and sleeping LL-TBE reference schemes and present our new algorithms. Then we prove their regret bounds and state their running times. Up to Section 4.4 we derive our results only for logarithmic loss, which allows us to use familiar concepts and results from probability

theory, like for example HMMs. In Section 4.5 we conclude by proving that any algorithm that satisfies certain weak conditions, in particular our generalisations of FS, directly generalises to an algorithm for arbitrary *mixable losses* with the appropriate regret bounds.

4.2 Notation: Prediction With Expert Advice

In this chapter we need do more extensive manipulation of segments of data, and corresponding predictions by EHMMs. This requires more elaborate notation than in the previous chapter, which we will proceed to introduce. Recall that the online learning setting of prediction with expert advice proceeds in rounds. In each round t, we first receive advice from a countable number of experts, which are indexed by $e \in \mathcal{E} \subseteq \mathbb{N}$. This advice comes in the form of an action $a_t^e \in \mathcal{A}$. Then we distill our own action $a_t \in \mathcal{A}$ from the expert advice. Finally, the actual outcome $x_t \in \mathcal{X}$ is observed, and everybody suffers loss as specified by a fixed loss function $\ell\colon \mathcal{A} \times \mathcal{X} \to [0, \infty]$. Thus, the performance of a sequence of actions $a_{1:T} = a_1, \ldots, a_T$ on data $x_{1:T} = x_1, \ldots, x_T$ is measured by the cumulative loss $\ell(a_{1:T}, x_{1:T}) = \sum_{t=1}^{T} \ell(a_t, x_t)$.

Log Loss We will initially present our results for *log(arithmic) loss* only, before generalising to a larger class of loss function in Section 4.5. For log loss the actions $\mathcal{A}$ are probability mass (or density) functions on $\mathcal{X}$ and $\ell(p, x) = -\log p(x)$ for any $p \in \mathcal{A}$, where log denotes the natural logarithm. Notice that minimising log loss is equivalent to maximising the predicted probability of outcome x. We write p_t^e for the prediction of expert e at time t and denote the predictions for all experts jointly by $p_t^{\mathcal{E}}$. Another important property of the log loss is the *chain rule*: interpreting any prediction $p_t(x_t)$ as the conditional density $p(x_t|x_{<t})$ of outcome x_t given all past outcomes $x_{<t} = x_1, \ldots, x_{t-1}$, we see that the cumulative log loss of a sequence of predictions

$$\sum_{t=1}^{T} -\log p_t(x_t) \;=\; -\log \prod_{t=1}^{T} p(x_t|x_{<t}) \;=\; -\log p(x_{1:T}) \tag{4.1}$$

equals the negative logarithm of the joint density $p(x_{1:T}) = \prod_{t=1}^{T} p(x_t|x_{<t})$ of all data $x_{1:T}$. Thus any lower bound on $p(x_{1:T})$ directly implies an upper bound on the cumulative loss of predictions $p_1, \ldots, p_T$ on data $x_{1:T}$.

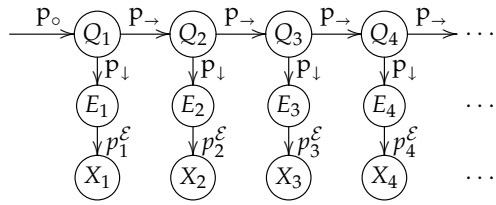


Figure 4.3: Bayesian network specification of an EHMM

Segments For $m \leq n$, we abbreviate the *segment* $\{m, \ldots, n\}$ to $m{:}n$. For any sequence $y_1, y_2, \ldots$ and any segment $\mathcal{C} = m{:}n$ we write $y_{\mathcal{C}}$ for the subsequence $y_m, \ldots, y_n$. For example, $x_{m:n} = x_m, \ldots, x_n$ and $p^{\mathcal{E}}_{1:T} = p^{\mathcal{E}}_1, \ldots, p^{\mathcal{E}}_T$. If all segments in a family $\mathbb{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots\}$ are pairwise disjoint and together cover $1{:}T$, then we call $\mathbb{C}$ a *segmentation* of $1{:}T$. We denote by $\langle e_{\mathcal{C}} \in \mathcal{E} \rangle_{\mathcal{C} \in \mathbb{C}}$ the labelling that assigns expert $e_{\mathcal{C}}$ to segment $\mathcal{C}$.

4.3 Expert Hidden Markov Models

EHMMs were introduced by Koolen and De Rooij [2008a] as a graphical and computational language to specify strategies for prediction with expert advice. EHMM diagrams directly represent the internal structure of the prediction strategy, facilitating the derivation of loss bounds. Moreover, there is a standard algorithm for sequential prediction, the *forward algorithm*, which greatly simplifies derivation of running time bounds.

In this chapter, we use EHMMs in two ways. On the input side, we use them to represent the learning experts whose predictions we want to combine. On the output side, we specify our own prediction strategies based on expert advice as EHMMs.

An *EHMM* $\mathfrak{H}$ is a probability distribution that is constructed according to the Bayesian network in Figure 4.3. It is used to sequentially predict outcomes X_1, X_2, …, which take values in outcome space $\mathcal{X}$, using advice from a set of experts $\mathcal{E}$. At each time t, the distribution of X_t depends on a hidden state Q_t, which determines mixing weights for the experts' predictions. Formally, the *production function* $\mathrm{p}_\downarrow$ determines the interpretation of a state: it maps any state $q_t \in \mathcal{E}$ to a distribution $\mathrm{p}_\downarrow^{q_t}$ on the identity E_t of the expert that should be used to predict X_t.

Then given $E_t = e$, the distribution of X_t is expert e's prediction p_t^e. It remains to define the distribution of the hidden states. The starting state Q_1 has *initial distribution* $\mathrm{p}_\circ$, and the state evolves according to the *transition function* $\mathrm{p}_\rightarrow$, which maps any state q_t to a distribution $\mathrm{p}_\rightarrow^{q_t}$ on its successor states.

An EHMM $\mathfrak{H}$ defines a prediction strategy as follows: after observing $x_{<t}$, predict the next outcome X_t using the marginal $\mathfrak{H}(X_t|x_{<t})$, which is a *mixture* of the experts' predictions $p_t^{\mathcal{E}}$.

We present four example EHMMs. The first three examples are suitable as input learning experts, which might be combined in the sleeping or freezing LL-TBE reference scheme. The fourth example represents FS as an EHMM, which will later be helpful when we compare it to our new generalisations.

Example 4.1 (Figure 4.1: Expert that Learns a Drifting Mean)**.** Here we formally define the EHMM $\mathsf{DM}[\theta]$ from the example in the introduction. Recall that the base experts predict according to standard normal distributions with fixed mean μ, which only takes integer values. Thus

$$p_t^\mu(x) := \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2}$$

for all $\mu \in \mathcal{E} := \mathbb{N} = \{0, 1, 2, \ldots\}$. In this EHMM it is sufficient to have a one-to-one correspondence between hidden states and experts, such that $Q_t = E_t$. This is expressed by $\mathcal{E} := \mathcal{E}$ and $\mathrm{p}_\downarrow := \mathbf{I}$, where $\mathbf{I}$ denotes the identity operator. The definition of $\mathsf{DM}[\theta]$ is completed by letting the initial distribution $\mathrm{p}_\circ$ be a point-mass on $\mu = 0$, and defining the transition function $\mathrm{p}_\rightarrow$ as in Figure 4.1: for any two states $\mu, \mu' \in \mathcal{E}$

$$\mathrm{p}_\rightarrow^\mu(\mu') := \begin{cases} \theta & \text{if } \mu' = \mu + 1, \\ 1 - \theta & \text{if } \mu' = \mu, \\ 0 & \text{otherwise.} \end{cases}$$

Example 4.2 (Bayes on base experts)**.** Consider the Bayesian mixture (also known as the exponentially weighted average predictor) of base experts $\mathcal{E}$ with prior w. We identify this prediction strategy with the following EHMM $\mathsf{B}[w]$, which makes the same predictions. As in the previous example, let $\mathcal{E} := \mathcal{E}$ and $\mathrm{p}_\downarrow := \mathbf{I}$, so that $Q_t = E_t$. This time, however, let $\mathrm{p}_\circ := w$ and $\mathrm{p}_\rightarrow := \mathbf{I}$. Despite its deceptive simplicity,

this EHMM *learns*: its marginal distribution of X_{t+1} given previous outcomes $x_{1:t}$ is a mixture of the base expert's predictions according to the Bayesian posterior.

Example 4.3 (Bayes on EHMMs)**.** Let $\mathcal{H} = \{\mathfrak{H}^1, \ldots, \mathfrak{H}^n\}$ be EHMMs with base experts $\mathcal{E}^1, \ldots, \mathcal{E}^n$, and let w be a prior on $\mathcal{H}$. Then, instead of treating $\mathfrak{H}^1, \ldots, \mathfrak{H}^n$ as black box predictors as in the previous example, their Bayesian mixture can also be expressed as a single EHMM $\mathsf{B}[w, \mathcal{H}]$ on the union of their base experts $\mathcal{E} := \bigcup_{i=1}^n \mathcal{E}^i$: assume without loss of generality that $\mathfrak{H}^1, \ldots, \mathfrak{H}^n$ have disjoint state spaces $\mathcal{E}^1, \ldots, \mathcal{E}^n$ and let $\mathcal{E} := \bigcup_{i=1}^n \mathcal{E}^i$. For any state $q \in \mathcal{E}^i$, let $\mathrm{p}^q_\downarrow$ equal $\mathrm{p}^{q,i}_\downarrow$, where $\mathrm{p}^i_\downarrow$ is the production function of $\mathfrak{H}^i$, so that all states keep their original interpretation. In addition let $\mathrm{p}_\circ(q) := w(i)\,\mathrm{p}^i_\circ(q)$, where $\mathrm{p}^i_\circ$ denotes the initial distribution of $\mathfrak{H}^i$. Finally, let $\mathrm{p}^q_\rightarrow(q')$ equal $\mathrm{p}^{q,i}_\rightarrow(q')$, the transition probability from q to q' for $\mathfrak{H}^i$ if $q, q' \in \mathcal{E}^i$ and let $\mathrm{p}^q_\rightarrow(q') := 0$ otherwise. Again, this EHMM *learns* which of the EHMMs in $\mathcal{H}$ is the best predictor.

Example 4.4 (Fixed-share)**.** The `Fixed-share` algorithm take a parameter α, called the *switching rate*. `Fixed-share` with prior distribution w on experts $\mathcal{E}$ and switching rate α can be represented as an EHMM $\mathsf{FS}[\alpha, w]$ as follows. As in the Bayesian mixture on base experts, let $\mathcal{E} := \mathcal{E}$ and $\mathrm{p}_\downarrow := \mathbf{I}$, so that $Q_t = E_t$, and let $\mathrm{p}_\circ := w$. Instead of the identity operator, however, use the transition function

$$\mathrm{p}_\rightarrow := (1-\alpha)\mathbf{I} + \alpha w \mathbf{1}^{\mathrm{T}},$$

where $\mathbf{1}^{\mathrm{T}}$ denotes the operator that sums the probability masses of all the hidden states. This transition function may be interpreted as follows: behave like the Bayesian mixture with probability $1-\alpha$, but with probability α take all the probability mass and redistribute it according to the prior w. (See also Figure 2.2 and the description of `Fixed-share` in Chapter 2.) Observe that for any probability distribution λ on states $\mathcal{E}$, we can compute $\mathrm{p}_\rightarrow \lambda = (1-\alpha)\lambda + \alpha w$ in constant time per state. We also note that in [Herbster and Warmuth, 1998] the prior w is always taken to be the uniform distribution, which gives the best worst-case regret bound.

4.3.1 Standard Fixed-share Loss Bound

To demonstrate the graphical derivation of loss bounds for EHMMs we now prove a regret bound for FS using its representation as an EHMM. The general technique is to give lower bounds on the transition function and the initial distribution. For simplicity the bound we show is slightly weaker than the standard regret bound [Herbster and Warmuth, 1998, Corollary 1]. (One could get the exact same bound by taking into account the remark in footnote 3 of [Koolen and De Rooij, 2008a], but this unnecessarily complicates the proof.)

Theorem 4.1. *Fix a prior w on experts $\mathcal{E}$ and a switching rate α. Then for any data $x_{1:T}$, expert predictions $p^{\mathcal{E}}_{1:T}$, reference segmentation $\mathbb{C}$ and assignment of experts to segments $\langle e_{\mathcal{C}} \in \mathcal{E} \rangle_{\mathcal{C}\in\mathbb{C}}$*

$$\ell\Big(\mathsf{FS}[\alpha, w], x_{1:T}\Big) \leq \underbrace{\sum_{\mathcal{C}\in\mathbb{C}} \ell(e_{\mathcal{C}}, x_{\mathcal{C}})}_{\text{S-TBE ref. scheme}} + \underbrace{(T-1)H(\alpha^*, \alpha)}_{\text{Switching}} + \underbrace{\sum_{\mathcal{C}\in\mathbb{C}} -\log w(e_{\mathcal{C}})}_{\text{Expert selection}},$$

where $H(\alpha, \beta) = -\alpha \log \beta - (1-\alpha)\log(1-\beta)$ and $\alpha^* = \frac{|\mathbb{C}|-1}{T-1}$.

Note that if w is the uniform distribution then $-\log w(e_{\mathcal{C}}) = \log|\mathcal{E}|$ for all $e_{\mathcal{C}}$. Then the difference with the standard bound in [Herbster and Warmuth, 1998] is $(|\mathbb{C}|-1)(\log|\mathcal{E}| - \log(|\mathcal{E}|-1))$, which is negligible.

Proof. Recall that $\mathsf{FS} \equiv \mathsf{FS}[\alpha, w]$ has transition function $\mathrm{p}_{\rightarrow} = (1-\alpha)\mathbf{I} + \alpha w \mathbf{1}^{\mathrm{T}}$. Therefore for any reference segmentation $\mathbb{C}$ the joint probability $\mathsf{FS}(x_{1:T})$ of any data sequence $x_{1:T}$ can be bounded from below by replacing transitions in FS *between* segments by $\alpha w \mathbf{1}^{\mathrm{T}}$, and those *within* the same segment by $(1-\alpha)\mathbf{I}$. The EHMM then degenerates into a sequence of independent Bayesian mixture EHMMs $\mathsf{B}[w]$ (see Example 4.2), one for each segment. Therefore

$$\mathsf{FS}(x_{1:T}) \geq \alpha^{|\mathbb{C}|-1}(1-\alpha)^{T-|\mathbb{C}|} \prod_{\mathcal{C}\in\mathbb{C}} \mathsf{B}[w](x_{\mathcal{C}}).$$

Similarly we can lower-bound the initial distribution of $\mathsf{B}[w]$ by a function that assigns weight $w(e_{\mathcal{C}})$ to the expert $e_{\mathcal{C}}$ selected for $\mathcal{C}$ in the reference segmentation and is 0 otherwise. It follows that $\mathsf{B}[w](x_{\mathcal{C}}) =$

$\sum_e w(e)p^e_{\mathcal{C}}(x_{\mathcal{C}}) \geq w(e_{\mathcal{C}})p^{e_{\mathcal{C}}}_{\mathcal{C}}(x_{\mathcal{C}})$, where $p^e_{\mathcal{C}}(x_{\mathcal{C}})$ denotes the joint probability of outcomes $x_{\mathcal{C}}$ according to the predictions of expert e. Hence by (4.1) we can conclude that

$$\begin{aligned}\ell\Big(\mathsf{FS}, x_{1:T}\Big) &= -\log \mathsf{FS}(x_{1:T}) \\ &\leq -\log \alpha^{|\mathbb{C}|-1}(1-\alpha)^{T-|\mathbb{C}|} + \sum_{\mathcal{C}\in\mathbb{C}} -\log p^{e_{\mathcal{C}}}_{\mathcal{C}}(x_{\mathcal{C}}) - \log w(e_{\mathcal{C}}) \\ &= (T-1)H(\alpha^*, \alpha) + \sum_{\mathcal{C}\in\mathbb{C}} \ell(e_{\mathcal{C}}, x_{\mathcal{C}}) + \sum_{\mathcal{C}\in\mathbb{C}} -\log w(e_{\mathcal{C}}),\end{aligned}$$

which completes the proof. □

4.4 Fixed-share for Learning Experts

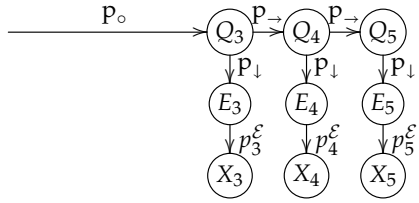
In this section we define the freezing and sleeping LL-TBE reference schemes for learning experts. Then, for each scheme, we provide our prediction strategy $\mathsf{FS}^{\mathsf{fr}}$ and $\mathsf{FS}^{\mathsf{sl}}$ and we prove that it achieves as small regret as FS.

4.4.1 LL-TBE and the Loss of an EHMM on a Segment

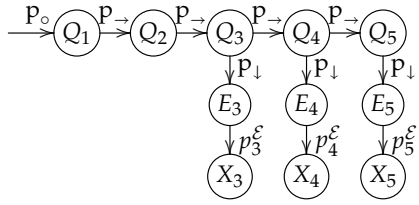
In order to state the loss of the freezing and sleeping LL-TBE reference schemes, we first define the loss of a single learning expert on a single segment. Then we define the loss of a whole segmentation.

Let $\mathfrak{H}$ be the EHMM for a learning expert with arbitrary base experts $\mathcal{E}$. Then the freezing and sleeping probability distributions $\mathfrak{H}^{\mathsf{fr}}_{i:j}$ and $\mathfrak{H}^{\mathsf{sl}}_{i:j}$ on segment $x_{i:j}$ are specified by the Bayesian networks of Figure 4.4. For freezing, the state at time i is simply initialised according to $\mathfrak{H}$'s initial distribution $\mathrm{p}_\circ$. For sleeping, we forward the initial distribution to time i by repeatedly applying the transition function $\mathrm{p}_\rightarrow$. Thus, the cumulative freezing and sleeping losses of $\mathfrak{H}$ on segment $x_{i:j}$ are given by $\ell(\mathfrak{H}^{\mathsf{fr}}_{i:j}, x_{i:j}) := -\log \mathfrak{H}^{\mathsf{fr}}_{i:j}(x_{i:j})$ and $\ell(\mathfrak{H}^{\mathsf{sl}}_{i:j}, x_{i:j}) := -\log \mathfrak{H}^{\mathsf{sl}}_{i:j}(x_{i:j})$. Note that we treat the base experts $\mathcal{E}$ as black boxes, so they may learn from the whole data.

Definition 4.1 (LL-TBE reference loss)**.** Fix data $x_{1:T}$ and a set of EHMMs $\mathcal{H}$. Let $\mathbb{C}$ be a segmentation of $1{:}T$ and let $\langle \mathfrak{H}_{\mathcal{C}} \in \mathcal{H}\rangle_{\mathcal{C}\in\mathbb{C}}$ be an assignment of experts to segments. Then the losses of the freez-



(a) Freezing: EHMM $\mathfrak{H}^{\mathsf{fr}}_{3:5}$



(b) Sleeping: EHMM $\mathfrak{H}^{\mathsf{sl}}_{3:5}$

Figure 4.4: Freezing and Sleeping EHMM $\mathfrak{H}$ on example segment $x_{3:5}$

ing and sleeping LL-TBE reference schemes are $\sum_{\mathcal{C}\in\mathbb{C}} \ell(\mathfrak{H}^{\mathsf{fr}}_{\mathcal{C}}, x_{\mathcal{C}})$ and $\sum_{\mathcal{C}\in\mathbb{C}} \ell(\mathfrak{H}^{\mathsf{sl}}_{\mathcal{C}}, x_{\mathcal{C}})$.

Note that selecting a learning expert on consecutive segments differs from selecting that expert on their union, since experts are reset between segments.

4.4.2 Main Result: Construction of the Freezing and Sleeping EHMMs

We now present the construction of EHMMs for the freezing and sleeping algorithms $\mathsf{FS}^{\mathsf{fr}}$ and $\mathsf{FS}^{\mathsf{sl}}$. Let $\mathcal{H}$ be a set of learning experts, each expert $\mathfrak{H} \in \mathcal{H}$ presented as an EHMM on basic experts $\mathcal{E}$. Let w be a prior on $\mathcal{H}$, and let α be a switching rate. We proceed in two steps. First construct the Bayesian EHMM $\mathfrak{B} = \mathsf{B}[w, \mathcal{H}]$ as in Example 4.3. Recall that $\mathfrak{B}$ learns which of the EHMMs in $\mathcal{H}$ predicts best. Second, construct the freezing EHMM $\mathsf{FS}^{\mathsf{fr}}[\alpha, \mathfrak{B}]$ or the sleeping EHMM[1] $\mathsf{FS}^{\mathsf{sl}}[\alpha, \mathfrak{B}]$ as shown in Figure 4.5. Note how, on a switch, both EHMMs reset the

[1]Strictly speaking, the Bayesian network in Figure 4.5b is not an EHMM, since the transition function depends on the time. Nevertheless, this time-dependency can be removed without any computational overhead using a process called *unfolding*, see [Koolen and De Rooij, 2008b].

entire state of $\mathfrak{B}$, which includes the states of experts in $\mathcal{H}$. In contrast, FS only resets its weighting on $\mathcal{H}$, but does not touch the internal state of the experts in $\mathcal{H}$.

4.4.3 Prediction Algorithms

To sequentially predict data using our prediction strategies $\mathsf{FS}^{\mathsf{fr}}$ and $\mathsf{FS}^{\mathsf{sl}}$, one needs to run the forward algorithm on their respective EHMMs. An explicit rendering of this process is included in Algorithm 4.1.

Algorithm 4.1 Explicit Forward Algorithm on FS^{v} for both Freezing and Sleeping ($\mathsf{v} \in \{\mathsf{fr}, \mathsf{sl}\}$)

1 Construct $\mathfrak{B} = \mathsf{B}[w, \mathcal{H}]$ with $\mathcal{E}$, $\mathrm{p}_\circ$,$\mathrm{p}_\downarrow$ and $\mathrm{p}_\rightarrow$ as in Example 4.3.
2 Initialisation: $\lambda \leftarrow \mathrm{p}_\circ$
3 **for** $t = 1, \ldots$ **do** ▷ *Invariant:* $\lambda(q) = \mathsf{FS}^{\mathsf{v}}[\alpha, \mathfrak{B}](Q_t = q | x_{<t})$
4 Receive expert advice $p_t^{\mathcal{E}}$.
5 Predict X_t using

$$\lambda(X_t) = \sum_{e \in \mathcal{E}, q \in \mathcal{E}} \lambda(q)\, \mathrm{p}_\rightarrow^q(e)\, p_t^e(X_t).$$

6 Observe $X_t = x_t$. Suffer loss $\ell\Big(\lambda(X_t), x_t\Big)$.
7 Loss update: $\lambda(q) \leftarrow \lambda(q, x_t)/\lambda(x_t)$, where

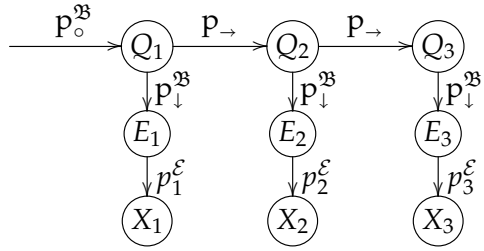
$$\lambda(q, x_t) = \sum_{e \in \mathcal{E}} \lambda(q)\, \mathrm{p}_\rightarrow^q(e)\, p_t^e(x_t).$$

8 State evolution:

$$\lambda \leftarrow \begin{cases} (1-\alpha)\, \mathrm{p}_\rightarrow \lambda + \alpha\, \mathrm{p}_\circ & \text{(Freezing)} \\ (1-\alpha)\, \mathrm{p}_\rightarrow \lambda + \alpha (\mathrm{p}_\rightarrow)^t\, \mathrm{p}_\circ & \text{(Sleeping)} \end{cases}$$

9 **end for**

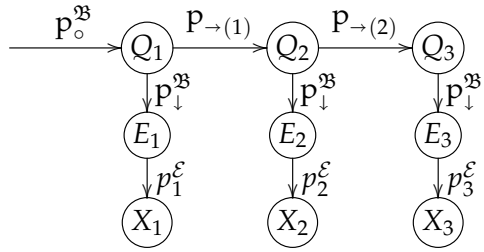
At any time t, the algorithm for $\mathsf{FS}^{\mathsf{sl}}$ only maintains non-zero weights on hidden states of the input learning experts that are reachable in *exactly* t steps from the starting states, just like the original FS algorithm. It therefore has the same running time. The algorithm for $\mathsf{FS}^{\mathsf{fr}}$, however, has to keep track of all states reachable in *at most* t steps. Consequently,



$$\mathrm{p}_\rightarrow \coloneqq (1-\alpha)\,\mathrm{p}_\rightarrow^{\mathfrak{B}} + \alpha\,\mathrm{p}_\circ^{\mathfrak{B}}\,\mathbf{1}^{\mathrm{T}}$$

Any switch reverts to $\mathrm{p}_\circ^{\mathfrak{B}}$, the initial distribution of $\mathfrak{B}$.

(a) EHMM $\mathsf{FS}^{\mathsf{fr}}[\alpha, \mathfrak{B}]$



$$\mathrm{p}_{\rightarrow(t)} \coloneqq (1-\alpha)\,\mathrm{p}_\rightarrow^{\mathfrak{B}} + \alpha(\mathrm{p}_\rightarrow^{\mathfrak{B}})^t\,\mathrm{p}_\circ^{\mathfrak{B}}\,\mathbf{1}^{\mathrm{T}}$$

The switch between time t and $t+1$ reverts to $(\mathrm{p}_\rightarrow^{\mathfrak{B}})^t\,\mathrm{p}_\circ^{\mathfrak{B}}$, the t^{th} evolution of the initial distribution of $\mathfrak{B}$.

(b) EHMM $\mathsf{FS}^{\mathsf{sl}}[\alpha, \mathfrak{B}]$

Figure 4.5: EHMMs for tracking the EHMM $\mathfrak{B}$ with switching rate α

in the worst case (over input EHMMs) it may be as slow as restarting expert copies (see Section 4.1.4). But if the input EHMMs have a finite number of hidden states, then its running time is of the same order as that of FS. And if the states (of the input EHMMs) that are reachable in exactly t steps are the same ones as the states reachable in at most t steps, which holds e.g. for the drifting-mean expert $\mathsf{DM}[\theta]$ from the introduction, then we also recover the efficiency of FS.

4.4.4 Loss Bound

Theorem 4.1 bounds the regret of FS compared to the S-TBE reference scheme by a "switching" and an "expert selection" term. We bound the regret of $\mathsf{FS}^{\mathsf{fr}}$ and $\mathsf{FS}^{\mathsf{sl}}$ compared to their LL-TBE reference scheme by the same two terms.

Theorem 4.2. *Fix a set of EHMMs $\mathcal{H}$ on basic experts $\mathcal{E}$, a prior w on $\mathcal{H}$, a switching rate α and $\mathsf{v} \in \{\mathsf{fr}, \mathsf{sl}\}$. Let $\mathfrak{B} = \mathsf{B}[w, \mathcal{H}]$. Then for any data $x_{1:T}$, expert predictions $p^{\mathcal{E}}_{1:T}$, reference segmentation $\mathbb{C}$ and assignment of experts to segments $\langle \mathfrak{H}_{\mathcal{C}} \in \mathcal{H} \rangle_{\mathcal{C} \in \mathbb{C}}$*

$$\ell\Big(\mathsf{FS}^{\mathsf{v}}[\alpha, \mathfrak{B}], x_{1:T}\Big) \;\leq\; \underbrace{\sum_{\mathcal{C} \in \mathbb{C}} \ell(\mathfrak{H}^{\mathsf{v}}_{\mathcal{C}}, x_{\mathcal{C}})}_{\text{LL-TBE ref. scheme}} + \underbrace{(T-1) H(\alpha^*, \alpha)}_{\text{Switching}} + \underbrace{\sum_{\mathcal{C} \in \mathbb{C}} -\log w(\mathfrak{H}_{\mathcal{C}})}_{\text{Expert selection}},$$

where $H(\alpha^*, \alpha)$ and $\alpha^* = \frac{|\mathbb{C}|-1}{T-1}$ are as in Theorem 4.1.

Proof. The proof proceeds like that of Theorem 4.1. Bounding transitions between segments from below by $\alpha\, \mathrm{p}^{\mathfrak{B}}_{\circ}\, \mathbf{1}^{\mathrm{T}}$ (freezing) or $\alpha (\mathrm{p}^{\mathfrak{B}}_{\rightarrow})^t\, \mathrm{p}^{\mathfrak{B}}_{\circ}\, \mathbf{1}^{\mathrm{T}}$ (sleeping), and transitions within each segment by $(1-\alpha)\, \mathrm{p}^{\mathfrak{B}}_{\rightarrow}$, we get

$$\mathsf{FS}^{\mathsf{v}}[\alpha, \mathfrak{B}] \geq \alpha^{|\mathbb{C}|-1} (1-\alpha)^{T-|\mathbb{C}|} \prod_{\mathcal{C} \in \mathbb{C}} \mathfrak{B}^{\mathsf{v}}_{\mathcal{C}}(x_{\mathcal{C}}), \tag{4.2}$$

where $\mathfrak{B}^{\mathsf{v}}_{\mathcal{C}}$ denotes the result of freezing or sleeping $\mathfrak{B}$ on segment $\mathcal{C} \in \mathbb{C}$ as in Figure 4.4. Observe that freezing and sleeping distribute over taking the Bayesian mixture: $\mathfrak{B}^{\mathsf{v}}_{\mathcal{C}} = \mathsf{B}[w, \mathcal{H}^{\mathsf{v}}_{\mathcal{C}}]$, where $\mathcal{H}^{\mathsf{v}}_{\mathcal{C}} := \{\mathfrak{H}^{\mathsf{v}}_{\mathcal{C}} \mid \mathfrak{H} \in \mathcal{H}\}$. As $\mathsf{B}[w, \mathcal{H}^{\mathsf{v}}_{\mathcal{C}}](x_{\mathcal{C}}) = \sum_{\mathfrak{H}} w(\mathfrak{H}) \mathfrak{H}^{\mathsf{v}}_{\mathcal{C}}(x_{\mathcal{C}}) \geq w(\mathfrak{H}_{\mathcal{C}}) \mathfrak{H}^{\mathsf{v}}_{\mathcal{C}}(x_{\mathcal{C}})$, the theorem follows from (4.1), like in the proof of Theorem 4.1. □

4.5 Other Loss Functions

We will now show how (our generalisations of) the `Fixed-share` algorithm for logarithmic loss can be directly translated into an algorithm with corresponding loss bound for any other mixable loss function. The same construction works for any logarithmic loss algorithm that predicts each outcome according to a mixture of the experts' predictions and whose predictions only depend on the experts' past losses on outcomes that actually occurred.

Mixability A loss function $\ell\colon \mathcal{A} \times \mathcal{X} \to [0,\infty]$ is called *η-mixable* for $\eta > 0$ if any distribution p on experts $\mathcal{E}$ can be mapped to a single action $\mathrm{Pred}(p) \in \mathcal{A}$ in a way that guarantees that

$$\ell\Big(\mathrm{Pred}(p), x\Big) \le -\tfrac{1}{\eta} \log \mathbf{E}_{e\sim p}\left[\exp\Big(-\eta\ell(a^e, x)\Big)\right] \tag{4.3}$$

for all outcomes $x \in \mathcal{X}$ and expert predictions $a^{\mathcal{E}}$. It is called *mixable* if it is η-mixable for some $\eta > 0$ [Cesa-Bianchi and Lugosi, 2006]. Mixability ensures that expert predictions for ℓ-loss can be mixed in essentially the same way as for log loss.

For example, logarithmic loss itself is 1-mixable. For $\mathcal{A} = [0,1]$ and $\mathcal{X} = \{0,1\}$ the *square loss* $\ell(a,x) := (a-x)^2$ is 2-mixable and the *Hellinger loss* $\ell(a,x) := ((\sqrt{1-x} - \sqrt{1-a})^2 + (\sqrt{x} - \sqrt{a}))/2$ is $\sqrt{2}$-mixable. A standard loss function that is not mixable is the *zero-one loss* for $\mathcal{A} = \mathcal{X} = \{0,1\}$, which is 0 if $a = x$ and 1 otherwise. Approaches for zero-one loss typically analyse expected loss under randomized actions, for which it can be approximated by mixable loss functions. [Haussler et al., 1998, Cesa-Bianchi and Lugosi, 2006]

The Benefits of Lying Given data $x_{1:t}$ and expert predictions $a^{\mathcal{E}}_{1:t}$, let $\ell^e_{1:t} := \ell(a^e_1, x_1), \ldots, \ell(a^e_t, x_t)$ denote the sequence of losses of expert e, and let $\ell^{\mathcal{E}}_{1:t}$ denote these losses jointly for all experts. From this point on we will write $\ell\ell$ instead of ℓ for the logarithmic loss.

Suppose alg is an algorithm for log loss that predicts each outcome x_t by mixing the experts' predictions $p^{\mathcal{E}}_t$ according to the distribution $p^{\mathsf{alg}}_t[x_{<t}, \ell\ell^{\mathcal{E}}_{<t}]$ on *experts*. The square-bracket expression indicates that p^{alg}_t may depend on the past outcomes $x_{<t} \equiv x_{1:t-1}$ and the losses $\ell\ell^{\mathcal{E}}_{<t}$ of the experts on these outcomes, but not on the experts' past or current

predictions in any other way. Following this convention, the algorithm predicts x_t using:

$$p_t^{\mathsf{alg}}[x_{<t}, \ell\ell^{\mathcal{E}}_{<t}](x_t) \;:=\; \sum_e p_t^{\mathsf{alg}}[x_{<t}, \ell\ell^{\mathcal{E}}_{<t}](e)\, p_t^e(x_t).$$

Now for any game with η-mixable loss ℓ and an equally large set of experts $\mathcal{E}$, we can derive from alg an algorithm alg^η_ℓ that predicts x_t according to

$$a_t^{\mathsf{alg}^\eta_\ell} \;:=\; \mathrm{Pred}\left(p_t^{\mathsf{alg}}[x_{<t}, \eta \cdot \ell^{\mathcal{E}}_{<t}]\right).$$

Note that alg^η_ℓ is lying to alg: while alg thinks it is playing a game for log loss in which experts have incurred log losses $\eta \cdot \ell^{\mathcal{E}}_{<t}$, in reality alg^η_ℓ is playing a game for loss ℓ and is feeding alg fake inputs and redirecting alg's outputs. Let us now analyse the loss of the derived algorithm alg^η_ℓ.

Theorem 4.3 (Other Loss Functions)**.** *Suppose* alg *is an algorithm for logarithmic loss that predicts according to* $p_t^{\mathsf{alg}}[x_{<t}, \ell\ell^{\mathcal{E}}_{<t}]$ *at each time* t*,* ℓ *is an* η*-mixable loss function, and* $f(x_{1:T}, \ell^{\mathcal{E}}_{1:T})$ *is an arbitrary function that maps outcomes and expert losses to real numbers. Then any log loss bound for* alg *of the form*

$$\ell\ell(\mathsf{alg}, x_{1:T}) \;\le\; f(x_{1:T}, \ell\ell^{\mathcal{E}}_{1:T}) \qquad \text{for all } p^{\mathcal{E}}_{1:T}, \tag{4.4}$$

directly implies a bound on the ℓ*-loss of* alg^η_ℓ*:*

$$\ell(\mathsf{alg}^\eta_\ell, x_{1:T}) \;\le\; \tfrac{1}{\eta} f(x_{1:T}, \eta \cdot \ell^{\mathcal{E}}_{1:T}) \qquad \text{for all } a^{\mathcal{E}}_{1:T}.$$

Proof. Construct a log loss game in which at any time t each expert e predicts according to a distribution p_t^e such that $p_t^e(x_t) = \exp(-\eta \ell_t^e)$ for the actual outcome x_t and p_t^e is arbitrary on other outcomes such that $\sum_{x_t} p_t^e(x_t) = 1$. By η-mixability (4.3) of ℓ we can relate the ℓ-loss of alg^η_ℓ to the log loss of alg:

$$\begin{aligned}
\ell(a^{\mathsf{alg}^\eta_\ell}_{1:T}, x_{1:T}) \;&=\; \sum_{t \in 1:T} \ell\Big(\mathrm{Pred}\left(p_t^{\mathsf{alg}}[x_{<t}, \eta \ell^{\mathcal{E}}_{<t}]\right), x_t\Big) \\
&\le\; \tfrac{1}{\eta} \sum_{t \in 1:T} -\log p_t^{\mathsf{alg}}[x_{<t}, \eta \ell^{\mathcal{E}}_{<t}](x_t) \\
&=\; \tfrac{1}{\eta} \ell\ell(\mathsf{alg}, x_{1:T}).
\end{aligned}$$

Combining with (4.4) completes the proof. □

Algorithms that satisfy the requirements of the theorem include `Follow-the-leader`, the `Mixing past posteriors` algorithm by Bousquet and Warmuth [2002] and any algorithm that can be represented as an EHMM, including `Fixed-share` and our generalisations, and the Bayesian mixture (Example 4.2). An algorithm that does not satisfy them is the `Last-step minimax` algorithm by Takimoto and Warmuth [2000], because it takes into account the experts' predictions on outcomes that do not occur.

In the literature it is common to construct algorithms for arbitrary mixable losses and point out their probabilistic interpretation for the special case of log loss [Haussler et al., 1998, Herbster and Warmuth, 1998, Bousquet and Warmuth, 2002]. Instead, we have proceeded the other way around: first we derived results for log loss and then we showed that they generalise to other losses. This allowed us to draw on concepts and results from probability theory like conditional probabilities, HMMs and the forward algorithm, without reproving them in a more general setting.

Theorem 4.3 generalises results by Vovk [1999], who shows that the most important loss bounds for Bayes with logarithmic loss can actually also be derived for arbitrary mixable losses. Our algorithm alg plays a role similar to his APA algorithm.

4.6 Conclusion

We revisited the tracking the best expert reference scheme (TBE), which asks for a strategy for prediction with expert advice that suffers small additional loss compared to the best expert per segment. This goal is natural when the characteristics of the data, and hence the best expert, are different between segments.

For learning experts, the standard interpretation of experts as black boxes implies training the experts on all data. We proposed a variation, adapted to learning experts, in which experts are only trained on the segment on which they are evaluated. Our scheme is able to exploit patterns in the data *per segment*, leading to smaller loss.

Although in general extending the standard `Fixed-share` algorithm to our setting will slow it down by a factor of T on T outcomes, we showed that no such slowdown is necessary if the learning experts can be represented as expert hidden Markov models (EHMMs). For arbi-

trary mixable losses we proved the loss bounds one would expect based on the loss bound for the original `Fixed-share` algorithm.

4.6.1 Discussion and Future Work

Learning the Switching Rate Like `Fixed-share`, our algorithms depend on a switching rate parameter α, which has to be fixed. Instead, one may want to tune α automatically based on the data. For FS this can be done efficiently, as described in the previous chapter. The same methods transfer directly to $\mathsf{FS}^{\mathsf{fr}}$ and $\mathsf{FS}^{\mathsf{sl}}$.

S-TBE vs LL-TBE We have discussed experts that learn only on their assigned segment. Perhaps surprisingly, this does *not always* increase performance. For example, we may have homogeneous data and experts that learn its global pattern at different rates. In such cases we clearly want to train each expert on all observations and, by switching at the right times, select the expert that has learned most until then. This scenario is analysed by Van Erven et al. [2008b], where experts are parameter estimators for a series of statistical models of increasing complexity.

Partitions instead of Segmentations Rather than split the data into segments as in the TBE reference scheme, one may wish to partition it arbitrarily into cells such that observations in the same cell need not be consecutive. Like `Fixed-share`, the corresponding algorithm [Bousquet and Warmuth, 2002] can be generalised to the LL-TBE setting without increasing its running time. In this case naively introducing copies of the experts for all possible partitions is infeasible: it would slow down the algorithm by an exponential factor 2^T on T outcomes.

Part II

MDL Convergence and Rényi Divergence

Chapter 5

MDL Convergence

In this chapter we investigate the scaling of the density code lengths that is required by Theorem 1.4 from Chapter 1. This scaling is equivalent to using the standard MDL estimator with density code lengths that satisfy a light-tails condition. It is found that if the scaling is simply removed, then MDL need not convergence at all. However, it is also shown that the light-tails condition can be weakened, and convergence also occurs for certain density code lengths that have heavy tails instead. The investigations in this chapter are preliminary, in the sense that the results that are obtained raise several natural follow-up questions which could not be addressed before finishing this thesis.

5.1 Introduction

In this chapter we return to the statistical terminology and notation of Chapter 1. Recall that, given a countable set of densities $\mathcal{M} = \{p_1, p_2, \ldots\}$, the *MDL estimator* maps any data $x^n \in \mathcal{X}^n$ to a density $\ddot{p}_n \in \mathcal{M}$ that achieves

$$\min_{p \in \mathcal{M}} L_n(p) - \log p(x^n), \tag{5.1}$$

where $L_n(p) = -\log \pi_n(p)$ is the *density code length* of p, and $\pi_n(p)$ is a (possibly incomplete) probability distribution on $\mathcal{M}$. As discussed in Chapter 1, the convention is to call π_n a *prior* even though it does not necessarily represent any prior beliefs. One may either take the density code lengths or the prior as primitive, as the other is easily derived. Throughout the chapter all logarithms will be natural logarithms, with base e.

Compression vs Convergence In Chapter 1, Theorem 1.4, it was shown that the MDL estimator converges to the true density at a rate determined by its description length of the data, but the result only holds if we scale the density code lengths $L_n(p)$ by a factor $\lambda > 1$, resulting in the *λ-MDL estimator*, which minimizes

$$\lambda L_n(p) - \log p(x^n). \tag{5.2}$$

Although from a frequentist point of view convergence may be all that is required, this result is very unsatisfying from a coding perspective, because there exists a code (with description lengths (5.1)) that allows for strictly better compression of the data. This is especially worrying in light of attempts to take the data compression interpretation as fundamental [Rissanen, 2007, Grünwald, 2007], which appeal for their clear operational interpretation that holds without probabilistic assumptions. The issue should also concern practitioners using Bayesian methods. As only standard MDL (with $\lambda = 1$) minimizes the Bayesian probability of error (see Chapter 1), being forced to take $\lambda > 1$ would mean acknowledging a fundamental problem with Bayesian methods as well. In fact, Zhang [2006] encounters exactly the same issue in analysing the convergence of the Bayesian posterior distribution and proposes to resolve it by modifying the standard Bayesian methods by introducing a similar $\lambda > 1$ parameter.

Light-Tails Condition In our presentation of the λ-MDL estimator, we have followed Zhang [2006] and Grünwald [2007]. It is well-known, however, that taking $\lambda > 1$ may equivalently be interpreted as using the *standard* MDL estimator with a condition on the density code lengths, which is called the *light-tails* condition [Barron and Cover, 1991]. To see this, consider alternative density code lengths

$$L'_n(p) = \lambda L_n(p) - c, \tag{5.3}$$

where c is a finite constant that does not depend on p or n. Clearly, using the standard MDL estimator with density code lengths L'_n is equivalent to using the λ-MDL estimator with density code lengths L_n. Taking for example $c = 0$, we see that the alternative density code lengths L'_n satisfy Kraft's inequality

$$\sum_p e^{-L'_n(p)} \leq 1$$

(see Theorem 1.1), and are therefore well-defined. Thus, the λ-MDL estimator can *always* be interpreted as the standard MDL estimator with alternative density code lengths. The light-tails condition comes in when we try to reverse the construction.

Suppose we use the standard MDL estimator with density code lengths L'_n. When do density code lengths L_n exist such that (5.3) is satisfied? Such L_n need to satisfy Kraft's inequality, which in terms of L'_n becomes:

$$\sum_p e^{-L'_n(p)/\lambda} \leq e^{c/\lambda}.$$

The condition simplifies when we express it in terms of the prior $\pi'_n(p) = e^{-L'_n(p)}$ and introduce $b = e^{c/\lambda}$:

$$\sum_p \pi'_n(p)^{1/\lambda} \leq b \qquad \text{for all } n. \tag{5.4}$$

For given $\lambda > 1$, we say that the density code lengths $L'_n(p) = -\log \pi'_n(p)$ satisfy the *light-tails* condition if (5.4) holds for some finite constant b, which does not depend on n. By (5.3), using standard MDL under the light-tails condition is equivalent to using λ-MDL (with different density code lengths), which is known to converge.

Gap with Consistency Theorem By Theorem 1.3, the standard MDL estimator is consistent for any choice of density code lengths that is independent of n, although the theorem says nothing about the rate at which it converges to the true density. For density code lengths that do not satisfy the light-tails condition, this behaviour is not explained by the convergence rate result Theorem 1.4. Thus, reexpressing scaling of the density code lengths as the light-tails condition, reveals a gap in current understanding of the behaviour of the MDL estimator.

Barron and Cover's Theorem In light of the previous discussion, it is worthwhile to investigate further. As a starting point for our investigations, we will take the convergence result that introduced the light-tails condition, by Barron and Cover [1991]. Let us describe the setting, which is similar to that of Theorem 1.4.

Suppose the data $X_1, \ldots, X_n$ are independent random variables, which are all distributed according to the same unknown density q, which need *not* be a member of the model $\mathcal{M}$. Barron and Cover

prove convergence of the MDL estimator in (squared) *Hellinger distance* $\mathrm{Hel}^2(q,p) = \int(\sqrt{q}-\sqrt{p})^2\,\mathrm{d}\mu$ at a rate determined by the *index of resolvability*

$$R_n(q) = \min_{p\in\mathcal{M}}\left\{\frac{1}{n}L_n(p) + D(q\|p)\right\}.$$

As discussed above, the light-tails condition implies that a factor λ is absorbed into the density code lengths. Taking this into account shows that the index of resolvability equals the right-hand side of (1.12) in Chapter 1.

Rather than in expectation, convergence is shown in probability. For any sequence of nonnegative random variables Y_n, convergence at positive rate R_n is denoted by $Y_n \lesssim R_n$ *in probability*. This means that the ratio Y_n/R_n is bounded in probability, i.e. for every $\varepsilon > 0$, there is a $c > 0$, such that $Q(Y_n/R_n > c) \leq \varepsilon$ for all large n.

It is further assumed that the density code lengths satisfy the *nondegeneracy* condition, which requires that there exists a constant $l > 0$ such that

$$L_n(p) \geq l \qquad \text{for all } p \in \mathcal{M} \text{ and all } n.$$

This condition is typically satisfied. For example, if the density code lengths are finite integers and $\mathcal{M}$ contains at least two densities, then we may take $l = 1$.

Theorem 5.1 ([Barron and Cover, 1991])**.** *Assume the density code lengths satisfy the light-tails condition and the nondegeneracy condition. If $R_n(q) \to 0$, then the standard MDL estimator converges to q in (squared) Hellinger distance, with rate bounded by the resolvability $R_n(q)$. That is,*

$$\mathrm{Hel}^2(q,\ddot{p}_n) \lesssim R_n(q) \qquad \textit{in probability.} \tag{5.5}$$

Theorem 5.1 has historically been important as a precursor to Theorem 1.4 and because it introduced the light-tails condition. Although convergence in probability is a rather weak mode of convergence, it is sufficient for our present investigations, since we are interested primarily in characterizing the *conditions* under which MDL converges. As will be seen below, the light-tails condition is not the weakest possible condition.

Outline We will first check whether the light-tails condition cannot just be dropped entirely. Unfortunately, by adapting an example from

[Zhang, 2006] it is found in Section 5.2 that MDL may not converge at all if no conditions on the density code lengths are imposed.

Then in Section 5.3 we investigate conditions that ensure convergence. We will be able to narrow down the set of problematic densities to a subset of the model with particular characteristics, which are expressed in terms of Rényi divergence. As the main result of this chapter, it will be shown that MDL still converges if this set has sufficiently small probability under the prior. It will be seen that for this to be the case, the light-tails condition is sufficient but not necessary. In particular, it is also sufficient if all density code lengths are equal (i.e., the prior is uniform), which is rather a *heavy-tails* condition. Sections 5.4 and 5.5 provide a technical discussion.

Our main result does not close the gap with the consistency theorem. We discuss this issue in Section 5.6. It is found that also under the conditions of the consistency theorem, the prior probability of the set of problematic densities goes to zero, albeit at an arbitrarily slow rate.

Section 5.7 provides some further discussion of the findings in this chapter. These findings raise multiple follow-up questions, which could not be addressed within the time available before finishing this thesis. In particular, it has not been tried to strengthen the results to stronger modes of convergence and connections to the large body of work on convergence of the Bayesian posterior distribution are not explored. In this sense, the present chapter is a preliminary study. Section 5.8 briefly reviews some of the remaining issues.

5.2 MDL Inconsistency Examples

We now present two examples in which with positive probability MDL selects a density $\ddot{p}_n$ that is very different from the true density q. In both examples MDL is inconsistent and does not converge. The examples show that the convergence rate results like Theorem 1.4 from Chapter 1 and our new result (Theorem 5.1) below, do not hold for standard MDL without imposing any further conditions.

5.2.1 Inconsistency for Arbitrary Partitions

For simplicity, let $\mathcal{X}$ be a countable sample space. Let Q denote the true distribution on $\mathcal{X}^n$ and let $\mathcal{P} = \{A_i\}_{i=1,2,\ldots}$ be an arbitrary partition of

$\mathcal{X}^n$ such that $Q(A_i) > 0$ for all i. Now let $\mathcal{M} = \{Q, P_1, P_2, \ldots\}$, where $P_i(x^n) = Q(x^n \mid A_i)$. Thus, the elements of $\mathcal{M}$ are all very similar to Q, except that they restrict attention to a specific element of the partition $\mathcal{P}$. Now let $\pi_n(Q) = 1/3$ and for any P_i let $\pi_n(P_i) = \frac{2}{3}Q(A_i)$. Let $L_n(P) = -\log \pi_n(P)$. Now we have for any sequence $x^n \in \mathcal{X}^n$ that

$$L_n(Q) - \log Q(x^n) = \log 3 - \log Q(x^n),$$

whereas for P_i such that $x^n \in A_i$

$$\begin{aligned} L_n(P_i) - \log P_i(x^n) &= -\log\left(\tfrac{2}{3}Q(A_i)\right) - \log Q(x^n \mid A_i) \\ &= \log \tfrac{3}{2} - \log Q(x^n). \end{aligned}$$

Thus on all data sequences (and therefore with probability one) MDL selects some P_i rather than the true distribution Q.

This example clearly illustrates that, although any individual P_i may be quite different from Q, the mixture of all P_i is very similar to Q:

$$\sum_i \pi_n(P_i) P_i = \frac{2}{3} Q.$$

This explains how densities with small prior probability, which by themselves are unlikely to be selected, can together still mislead the MDL estimator.

5.2.2 Inconsistency for Sample Size Dependent Prior

The construction of the previous example depends on the sample size n and the data are not independent and identically distributed (i.i.d.) under the distributions $P_i \in \mathcal{M}$. This raises the question of whether MDL can still be inconsistent when $\mathcal{M}$ contains only i.i.d. distributions. The present example shows that this is the case by embedding an example for the multinomial model from [Zhang, 2006] in a continuous setting. We have to modify Zhang's example, because in his version the sample space and the distributions in the model all depend on n. In our version the only dependence on n is through the prior. This last dependence cannot be removed, because otherwise the MDL estimator would be consistent by Theorem 1.3.

Let $\mathcal{X} = [0, 1]$ be the unit interval and let $X^n = X_1, \ldots, X_n$ be i.i.d. random variables, taking values in $\mathcal{X}$ according to the uniform density

q. Let $\mathcal{M}' = \{p_i^k \mid k = 2, 4, 6, \ldots, i = 1, \ldots, \binom{k}{k/2}\}$ denote all histograms on $\mathcal{X}$ with an even number of bins k of equal width that put density 2 on exactly half of the bins and density 0 on the other bins. Let $\mathcal{M} = \mathcal{M}' \cup \{q\}$. All densities are extended to n outcomes by taking products: $p(x^n) = \prod_{i=1}^n p(x_i)$.

For any sample size n, let $m \gg n$ be sufficiently large that $(m-n)^n/m^n \geq \frac{1}{2}$. Then define $L_n(p) = -\log \pi_n(p)$ and the sample size dependent prior π_n as

$$\pi_n(q) = \frac{1}{4},$$
$$\pi_n(p_i^k) = \begin{cases} \frac{1}{4}w(p_i^k) + 2^{-n-1} & \text{if } k = 2m \text{ and } i \leq 2^n, \\ \frac{1}{4}w(p_i^k) & \text{otherwise,} \end{cases}$$

where w is an arbitrary positive prior on the submodel $\mathcal{M}'$ that does not depend on n. It is necessary to have π_n depend on n in order to make the example work. Otherwise, by Theorem 1.3, MDL would almost surely select q for all sufficiently large n. Note however that $L_n(p)/n \to 0$ for all $p \in \mathcal{M}$.

If $p_i^{2m}(x^n) > 0$ for $i \leq 2^n$, then

$$L_n(p_i^{2m}) - \log p_i^{2m}(x^n) \leq -\log 2^{-n-1} - \log 2^n = \log 2,$$

whereas

$$L_n(q) - \log q(x^n) = \log 4 > \log 2,$$

and consequently MDL prefers p_i^{2m} over q. It follows that

$$Q(\ddot{p}_n \neq q) \geq 1 - Q\left(\forall i \leq 2^n : p_i^{2m}(X^n) = 0\right).$$

Now consider the set $\{p_i^{2m} \mid i = 1, \ldots, \binom{2m}{m}\}$ of all histograms with $2m$ bins. By symmetry it does not matter how we order the elements of this set, which determines which elements receive the extra prior mass 2^{-n-1}. Consequently, we could equally well have sampled the elements with extra prior randomly without replacement after sampling the data. But then the probability of the event $A_n = \{\forall i \leq 2^n : p_i^{2m}(X^n) = 0\}$ would increase if we would sample *with replacement*. Let $\Pr(A_n \mid X^n)$ denote the conditional probability of A_n given X^n when the elements

with extra prior mass are sampled uniformly with replacement. Then, like in Zhang's example,

$$Q\Big(\forall i \le 2^n : p_i^{2m}(X^n) = 0\Big) \le \mathbf{E}_{X^n} \Pr\Big(\forall i \le 2^n : p_i^{2m}(X^n) = 0 \mid X^n\Big)$$
$$= \mathbf{E}_{X^n} \Pr\Big(p_1^{2m}(X^n) = 0 \mid X^n\Big)^{2^n} = \mathbf{E}_{X^n}\left(1 - \frac{\binom{2m-|X^n|}{m-|X^n|}}{\binom{2m}{m}}\right)^{2^n}$$
$$= \mathbf{E}_{X^n}\left(1 - \prod_{i=0}^{|X^n|-1} \frac{m-i}{2m-i}\right)^{2^n} \le \left(1 - \left(\frac{m-n}{2m}\right)^n\right)^{2^n} \le e^{-1/2},$$

where $|X^n| \le n$ denotes the number of different bins (out of $2m$ bins) that contain at least one outcome from $X_1, \ldots, X_n$, and the last inequality follows by $(m-n)^n/m^n \ge \frac{1}{2}$ and $1 + t \le e^t$ for all t.

Like in Zhang's original example, all elements of the submodel $\mathcal{M}'$ have the same, strictly positive divergence from q in all the usual divergence measures, like Rényi divergence, Kullback-Leibler divergence and Hellinger distance. Thus, for all n, with probability at least $1 - e^{-1/2}$ the density selected by MDL is at a fixed distance (independent of n) from q. It follows that the MDL estimates do not converge.

5.3 Weakening the Light-Tails Condition

In the previous section we have seen how the MDL estimator may be inconsistent if the model, the prior and the sample size are chosen adversarially. Let us try to characterize the conditions under which this may happen. Our characterization involves the *Rényi divergence* $D_\alpha(p_1\|p_2) = \frac{1}{\alpha-1}\log\int p_1^\alpha p_2^{1-\alpha}\,\mathrm{d}\mu$ of *order* $\alpha \ne 1$ of p_1 from p_2 (see Chapter 6), which is nondecreasing in α. As long as it is finite, Rényi divergence is also continuous in α and tends to the Kullback-Leibler divergence $D(p_1\|p_2) = \int p_1 \log(p_1/p_2)\,\mathrm{d}\mu$ as α tends to 1, which is therefore how it is defined for $\alpha = 1$.

Let the data be sampled i.i.d. according to q, which does not have to be an element of $\mathcal{M}$, and let $\tilde{q}_n \in \mathcal{M}$ realize the index of resolvability, i.e.

$$\frac{1}{n}L_n(\tilde{q}_n) + D(q\|\tilde{q}_n) = R_n(q).$$

We find that if MDL selects a bad density at all, then this density has to be a member of the set

$$A_n = \left\{ p \in \mathcal{M} \middle| c_1 D_\alpha(p\|\tilde{q}_n) < \frac{L_n(p) - L_n(\tilde{q}_n)}{n} < c_2 D_\beta(p\|\tilde{q}_n) \right\} \quad (5.6)$$

for orders $0 < \alpha < 1 < \beta$ and constants $0 < c_1 < 1 < c_2$ which will be discussed below. Inconsistency is avoided, however, if the prior probability of the elements of A_n decreases exponentially with their divergence from $\tilde{q}_n$, as in the following condition:

Condition 5.1. There exist constants $b \geq 0$ and $k > 0$ such that for all n, whatever the identity of $\tilde{q}_n \in \mathcal{M}$ is,

$$\pi_n(\mathcal{E} \cap A_n) \leq b e^{-kn\varepsilon} \quad \text{for all } \varepsilon > 0, \quad (5.7)$$

where $\mathcal{E} = \{p \in \mathcal{M} \mid D_\alpha(p\|\tilde{q}_n) \geq \varepsilon\}$, and A_n is as in (5.6) with some choices of $0 < \alpha < 1 < \beta$ and $0 < c_1 < 1 < c_2$.

Note that since $\tilde{q}_n$ depends on the true density q, which is unknown, the condition requires that we check (5.7) for all possibilities.

We will show that Condition 5.1 is weaker than the light-tails condition. Nevertheless, MDL still converges at a rate determined by the index of resolvability if the light-tails condition is replaced by Condition 5.1:

Theorem 5.2. *Assume the density code lengths satisfy Condition 5.1 and the nondegeneracy condition. If $R_n(q) \to 0$, then the standard MDL estimator converges to q in (squared) Hellinger distance, with rate bounded by the resolvability $R_n(q)$. That is,*

$$\mathrm{Hel}^2(q, \ddot{p}_n) \lesssim R_n(q) \qquad \textit{in probability.} \quad (5.8)$$

We note that any fixed choice of α, β, c_1 and c_2 in Condition 5.1 is sufficient for the theorem to hold, but as any of these goes to 1, the implicit constants of the theorem deteriorate until it becomes vacuous (see Lemma 5.2 below). The proof of the theorem is given in Section 5.5. In future work we would hope to strengthen this result to convergence in expectation for finite samples, like in Theorem 1.4. This has not yet been attempted.

5.3.1 Satisfying Condition 5.1

Although at first sight Condition 5.1 might appear complicated, there are at least two important cases in which it is easy to verify. The first arises when all density code lengths $L_n(p)$ are equal, because then non-negativity of Rényi divergence implies that A_n is empty. Note that this corresponds to a uniform prior π_n, which is important for its data compression properties, as discussed in Section 1.3.5. In fact, the same reasoning still applies if the prior varies sufficiently slowly relative to Rényi divergence:

Proposition 5.1 (Uniformish Prior)**.** *If*

$$L_n(p) - L_n(p') \leq c_1 n D_\alpha(p\|p')$$

for all $p, p' \in \mathcal{M}$, then $A_n = \emptyset$ for any β and c_2, and Condition 5.1 is satisfied with $b = 0$.

This proposition generalises the following observation, which could already be made based on previous results: if all density code lengths are exactly equal, then the λ-MDL estimator and the ordinary MDL estimator coincide. Hence in this special case convergence of λ-MDL implies convergence of the ordinary MDL estimator at the same rate. However, this ad-hoc observation breaks down as soon as one allows minor variations in the density code lengths, and one may wonder what happens to ordinary MDL in such cases. Proposition 5.1 then shows that it continues to converge.

There is a second case in which Condition 5.1 is easy to verify. This is when the light-tails condition is satisfied:

Lemma 5.1. *Suppose, for $0 < \alpha < 1$, there exists a constant $b < \infty$ such that*

$$\sum_p \pi_n(p)^{1-\alpha} \leq b, \qquad \text{for all } n. \tag{5.9}$$

Then Condition 5.1 is satisfied with the same constant b and $k = \alpha c_1$.

Proof.

$$\begin{aligned}\pi_n(\mathcal{E} \cap A_n) &= \sum_{p \in \mathcal{E} \cap A_n} \pi_n(p) \leq \sum_{p \in \mathcal{E} \cap A_n} \pi_n(p)^{1-\alpha} e^{-\alpha(L_n(p) - L_n(\tilde{q}_n))} \\ &\leq \sum_{p \in \mathcal{E} \cap A_n} \pi_n(p)^{1-\alpha} e^{-\alpha c_1 n D_\alpha(p \| \tilde{q}_n)} \\ &\leq \sum_{p \in \mathcal{E} \cap A_n} \pi_n(p)^{1-\alpha} e^{-\alpha c_1 n \varepsilon} \leq b e^{-\alpha c_1 n \varepsilon}. \qquad \square\end{aligned}$$

Thus, the light-tails condition implies Condition 5.1. The other way around, however, this is not the case. Consider, for example, the Bernoulli example (Example 1.4) from Chapter 1, in which $m \approx \sqrt{n}$ densities were all assigned the same code length $\log m$. This satisfies Condition 5.1 (by Proposition 5.1), but the light-tails condition does not hold:

$$\sum_p \left(\frac{1}{m}\right)^{1-\alpha} = m \left(\frac{1}{m}\right)^{1-\alpha} \approx n^{1/2} n^{-(1-\alpha)/2} = n^{\alpha/2} \to \infty.$$

This shows that the light-tails condition is strictly stronger than Condition 5.1. Note that, surprisingly, to satisfy the uniformity requirements of Proposition 5.1, the prior must have heavy tails, instead of light tails.

Remark 5.1. It is shown in Chapter 6 that

$$\mathrm{Hel}^2(p_1, p_2) \leq D_{1/2}(p_1 \| p_2) \leq D_2(p_1 \| p_2) \leq \chi^2(p_1, p_2),$$

where $\chi^2(p_1, p_2) = \int (p_1 - p_2)^2 / p_2 \, \mathrm{d}\mu$ denotes the χ^2-distance. For $\alpha = 1/2$ and $\beta = 2$, Condition 5.1 is therefore implied if A_n is replaced by the larger set

$$A_n' = \left\{ p \in \mathcal{M} \,\middle|\, c_1 \,\mathrm{Hel}^2(p, \tilde{q}_n) < \frac{L_n(p) - L_n(\tilde{q}_n)}{n} < c_2 \chi^2(p, \tilde{q}_n) \right\}.$$

This, however, is a significantly stronger condition. For example, if p and $\tilde{q}_n$ are mutually singular, then $D_{1/2}(p \| \tilde{q}_n) = \infty$, but $\mathrm{Hel}^2(p, \tilde{q}_n) = 2$, so that for a large range of density code lengths

$$\mathrm{Hel}^2(p, \tilde{q}_n) < \frac{L_n(p) - L_n(\tilde{q}_n)}{n} \leq D_{1/2}(p \| \tilde{q}_n).$$

5.4 Chernoff Bound

The results by Barron and Cover [1991] all depend on an inequality that is essentially *Chernoff's bound* [Cover and Thomas, 1991]

$$\tilde{Q}_n(Z \geq a) \leq e^{-\gamma a} \mathbf{E}_{\tilde{Q}_n}[e^{\gamma Z}] \qquad (\gamma \geq 0),$$

applied to the random variable $Z = \log p(X^n)/\tilde{q}_n(X^n)$ with $a = L_n(p) - L_n(\tilde{q}_n)$. The bound leaves open the choice of γ, which ideally should be tuned to make the bound as tight as possible. When Barron and Cover apply the Chernoff bound, they use the same choice of γ for all p and $\tilde{q}_n$. In particular, if their light-tails condition is satisfied with $\alpha = 1/2$, they take $\gamma = 1/2$. The proof of the following lemma, which is the key to proving Theorem 5.2, refines this approach by letting γ depend on p and $\tilde{q}_n$. The choice of γ is discussed after the proof.

Lemma 5.2. *Let $\tilde{q}_n \in \mathcal{M}$ and $\varepsilon > 0$. Then for any orders $0 < \alpha < 1 < \beta$ and any constants $0 < c_1 < 1 < c_2$*

$$\tilde{Q}_n\left(\ddot{p}_n \in \mathcal{E}\right) \leq \sum_{p \in \mathcal{E} \setminus A_n} \frac{\pi_n(p)}{\pi_n(\tilde{q}_n)} e^{-c' n D_\alpha(p\|\tilde{q}_n)} + \frac{\pi_n(\mathcal{E} \cap A_n)}{\pi_n(\tilde{q}_n)}, \tag{5.10}$$

where $\mathcal{E} = \{p \in \mathcal{M} \mid D_\alpha(p\|\tilde{q}_n) \geq \varepsilon\}$ and A_n are as in Condition 5.1, and $c' = \min\{(1-c_1)(1-\alpha), (c_2-1)(\beta-1)\}$.

Proof. The event $\ddot{p}_n \in \mathcal{E}$ only occurs if there exists some $p \in \mathcal{E}$ such that

$$L_n(p) - \log p(X^n) \leq L_n(\tilde{q}_n) - \log \tilde{q}_n(X^n). \tag{5.11}$$

For arbitrary $p \in \mathcal{E}$, let B_p denote the event (5.11). Then Chernoff's bound, applied as discussed above, implies that for any $\gamma \geq 0$

$$\begin{aligned}\tilde{Q}_n(B_p) &\leq \left(\frac{\pi_n(p)}{\pi_n(\tilde{q}_n)}\right)^\gamma \mathbf{E}_{\tilde{Q}_n}\left(\frac{p(X^n)}{\tilde{q}_n(X^n)}\right)^\gamma \\ &\leq \frac{\pi_n(p)}{\pi_n(\tilde{q}_n)} e^{(\gamma-1)(nD_\gamma(p\|\tilde{q}_n) - (L_n(p) - L_n(\tilde{q}_n)))},\end{aligned}$$

where the last inequality follows from additivity of Rényi divergence and holds with equality unless $\gamma \geq 1$ and $P \not\ll \tilde{Q}_n$, in which case $D_\gamma(p\|\tilde{q}_n) = \infty$. The lemma now follows by the union bound,

$$\tilde{Q}_n\left(\ddot{p}_n \in \mathcal{E}\right) \leq \sum_{p \in \mathcal{E}} \tilde{Q}_n(B_p),$$

and the following choices for γ: for $p \in A_n$, take $\gamma = 1$ to get

$$\tilde{Q}_n(B_p) \leq \frac{\pi_n(p)}{\pi_n(\tilde{q}_n)};$$

for p such that $L_n(p) - L_n(\tilde{q}_n) \leq c_1 n D_\alpha(p\|\tilde{q}_n)$, take $\gamma = \alpha$ to get

$$\tilde{Q}_n(B_p) \leq \frac{\pi_n(p)}{\pi_n(\tilde{q}_n)} e^{-(1-\alpha)(1-c_1)nD_\alpha(p\|\tilde{q}_n)};$$

and for p such that $L_n(p) - L_n(\tilde{q}_n) \geq c_2 n D_\beta(p\|\tilde{q}_n)$, take $\gamma = \beta$ to get

$$\tilde{Q}_n(B_p) \leq \frac{\pi_n(p)}{\pi_n(\tilde{q}_n)} e^{-(\beta-1)(c_2-1)nD_\beta(p\|\tilde{q}_n)} \leq \frac{\pi_n(p)}{\pi_n(\tilde{q}_n)} e^{-(\beta-1)(c_2-1)nD_\alpha(p\|\tilde{q}_n)},$$

where the second inequality follows from the fact that Rényi divergence is nondecreasing in its order. □

The preceding proof applies Chernoff's bound with different choices of γ, depending on p and $\tilde{q}_n$. These choices can be motivated as follows. (See also Section 6.6.1 in the next chapter, which provides a related discussion.) Under regularity conditions, Grünwald [2007, p. 648] shows that $(1-\gamma)D_\gamma(p\|\tilde{q}_n)$ is strictly concave in γ and

$$\frac{\mathrm{d}}{\mathrm{d}\gamma}(1-\gamma)D_\gamma(p\|\tilde{q}_n) = D(p_\gamma\|p) - D(p_\gamma\|\tilde{q}_n),$$

where $p_\gamma = p^\gamma \tilde{q}_n^{1-\gamma} / \int p^\gamma \tilde{q}_n^{1-\gamma}\,\mathrm{d}\mu$. (Note that $p_0 = \tilde{q}_n$ and $p_1 = p$.) The exponent we get from Chernoff's bound,

$$(\gamma - 1)\Big(nD_\gamma(p\|\tilde{q}_n) - (L_n(p) - L_n(\tilde{q}_n))\Big), \tag{5.12}$$

is therefore strictly convex in γ and minimal at γ^* such that

$$n\Big(D(p_{\gamma^*}\|\tilde{q}_n) - D(p_{\gamma^*}\|p)\Big) = L_n(p) - L_n(\tilde{q}_n).$$

Suppose

$$D(p\|\tilde{q}_n) \approx \frac{L_n(p) - L_n(\tilde{q}_n)}{n}.$$

Then $\gamma^* \approx 1$ and the exponent in (5.12) is approximately 0. As Rényi divergence is nondecreasing and, if finite, also continuous in its order,

this is the case for densities in A_n (assuming that the parameters α and β in the definition of A_n are close to 1). Densities not in A_n we split into two categories: those for which $\gamma^* < 1$ and those for which $\gamma^* > 1$. In the first case we apply Chernoff's bound with $\gamma = \alpha$, and in the second case with $\gamma = \beta$. Although this argument shows that we could get an even better bound for $p \notin A_n$ by tweaking γ even further, these are not the p that prevent MDL from converging, so this optimization is unnecessary for our present purposes. Only $p \in A_n$ lead to problems, and for these we already use (almost) the optimal γ, so there is no further room for improvement using Chernoff's bound.

5.5 Proof of Theorem 5.2

Our proof of Theorem 5.2 closely parallels the proof of Theorem 5.1 by Barron and Cover [1991], except that their application of Chernoff's bound is replaced by the following lemma:

Lemma 5.3. *Suppose Condition 5.1 is satisfied with constants b, k and α, β, c_1, c_2. Then for any $\tilde{q}_n \in \mathcal{M}$ and $\varepsilon > 0$*

$$\tilde{Q}_n\left(D_\alpha(\ddot{p}_n \| \tilde{q}_n) \geq \varepsilon\right) \leq (1+b)e^{-an\varepsilon + L_n(\tilde{q}_n)}, \tag{5.13}$$

where $a = \min\{k, (1-c_1)(1-\alpha), (c_2-1)(\beta-1)\}$.

Proof. By Lemma 5.2

$$\tilde{Q}_n\left(D_\alpha(\ddot{p}_n \| \tilde{q}_n) \geq \varepsilon\right) \leq \sum_{p \in \mathcal{E} \setminus A_n} \frac{\pi_n(p)}{\pi_n(\tilde{q}_n)} e^{-c' n D_\alpha(p\|\tilde{q}_n)} + \frac{\pi_n(\mathcal{E} \cap A_n)}{\pi_n(\tilde{q}_n)}.$$

As $D_\alpha(p\|\tilde{q}_n) \geq \varepsilon$ for all $p \in \mathcal{E}$, the first term on the right-hand side may be bounded by

$$\begin{aligned}\sum_{p \in \mathcal{E} \setminus A_n} \frac{\pi_n(p)}{\pi_n(\tilde{q}_n)} e^{-c' n D_\alpha(p\|\tilde{q}_n)} &\leq \sum_{p \in \mathcal{E} \setminus A_n} \frac{\pi_n(p)}{\pi_n(\tilde{q}_n)} e^{-an\varepsilon} \\ &= e^{-an\varepsilon + L_n(\tilde{q}_n)} \pi_n\left(\mathcal{E} \setminus A_n\right) \leq e^{-an\varepsilon + L_n(\tilde{q}_n)}.\end{aligned}$$

The lemma follows by combining these bounds with the bound from Condition 5.1. □

Note that Lemma 5.3 is where Condition 5.1 comes in: it ensures that the sum over Rényi divergences in Lemma 5.2 dominates the prior term, which gives an exponentially small probability in ε that the MDL estimator diverges more than ε from $\tilde{q}_n$.

To deal with densities q outside of $\mathcal{M}$, we use the following lemma by Barron and Cover, which allows us to change measures from q to $\tilde{q}_n$.

Lemma 5.4 ([Barron and Cover, 1991])**.** *Let q and $\tilde{q}_n$ be densities on $\mathcal{X}$ and let $X^n = X_1, \ldots, X_n$ be independent random variables with density q or $\tilde{q}_n$. Then*

$$Q(X^n \in B) \leq \tilde{Q}_n(X^n \in B)e^{nr} + \frac{D(q\|\tilde{q}_n)}{r} + \frac{1}{enr}$$

for any measurable event $B \subseteq \mathcal{X}^n$ and $r > 0$.

The remainder of the proof of Theorem 5.2 is very similar to the proof of Barron and Cover.

Proof of Theorem 5.2. Let b, k, α, β, c_1 and c_2 be constants that satisfy Condition 5.1, and let $0 < a' < 1$ be a new constant to be specified later. For $c > 1/a' > 1$, let

$$B_n = \left\{ \mathrm{Hel}^2(q, \ddot{p}_n) > 4cR_n(q) \right\}.$$

For arbitrary $\varepsilon > 0$, we show that $Q(B_n) \leq \varepsilon$ for all n if c is sufficiently large. To this end, we apply Lemma 5.4 with $r = (a'c - 1)R_n(q)/2$, and use that $R_n(q) \geq D(q\|\tilde{q}_n)$ and $nR_n(q) \geq L_n(\tilde{q}_n) \geq l$ (by the nondegeneracy assumption), which yields

$$Q(B_n) \leq \tilde{Q}_n(B_n)e^{(a'c-1)nR_n(q)/2} + \frac{2 + 2/(el)}{a'c - 1}. \tag{5.14}$$

We proceed to bound $\tilde{Q}_n(B_n)$. Using the triangle inequality

$$\mathrm{Hel}(q, \ddot{p}_n) \leq \mathrm{Hel}(q, \tilde{q}_n) + \mathrm{Hel}(\tilde{q}_n, \ddot{p}_n)$$

and the bound $\mathrm{Hel}^2(q, \tilde{q}_n) \leq D_{1/2}(q\|\tilde{q}_n) \leq D(q\|\tilde{q}_n) \leq R_n(q)$, we find that on B_n

$$\begin{aligned}\mathrm{Hel}(\tilde{q}_n, \ddot{p}_n) &\geq \mathrm{Hel}(q, \ddot{p}_n) - \mathrm{Hel}(q, \tilde{q}_n) \\ &> (2\sqrt{c} - 1)\sqrt{R_n(q)} > \sqrt{cR_n(q)}.\end{aligned}$$

Consequently, by Theorem 6.17 in Chapter 6 and symmetry of Hellinger distance,

$$\begin{aligned} mD_\alpha(\ddot{p}_n\|\tilde{q}_n) &\geq D_{1/2}(\ddot{p}_n\|\tilde{q}_n) \\ &\geq \mathrm{Hel}^2(\ddot{p}_n,\tilde{q}_n) = \mathrm{Hel}^2(\tilde{q}_n,\ddot{p}_n) > cR_n(q) \end{aligned}$$

on B_n, for $m = \max\{1,(1-\alpha)/\alpha\}$. It follows that B_n is a subset of the event

$$\tilde{B}_n = \left\{D_\alpha(\ddot{p}_n\|\tilde{q}_n) > cR_n(q)/m\right\},$$

and therefore that $\tilde{Q}_n(B_n) \leq \tilde{Q}_n(\tilde{B}_n)$. Hence by Lemma 5.3

$$\tilde{Q}_n(B_n) \leq \tilde{Q}_n(\tilde{B}_n) \leq (1+b)e^{-a'cnR_n(q)+L_n(\tilde{q})} \leq b'e^{-(a'c-1)nR_n(q)},$$

where $b' = 1+b$ and we now specify that $a' = a/m$. Plugging this into (5.14) gives

$$\begin{aligned} Q(B_n) &\leq b'e^{-(a'c-1)nR_n(q)/2} + \frac{2+2/(el)}{a'c-1} \\ &\leq b'e^{-(a'c-1)l/2} + \frac{2+2/(el)}{a'c-1}, \end{aligned}$$

which does not exceed ε for sufficiently large c, as required. □

5.6 The Gap with Consistency

We have seen that the MDL estimator converges at a rate determined by the index of resolvability if the density code lengths satisfy Condition 5.1, which comes in to ensure that the prior probability of A_n converges to zero at a sufficiently fast rate. What happens if, instead of Condition 5.1, we impose the conditions of the consistency theorem (Theorem 1.3)?

The two conditions of the consistency theorem are that $q \in \mathcal{M}$ and that the density code lengths do not vary with n. Consistency then implies that

$$\mathrm{Hel}^2(q,\ddot{p}_n) = 0 \quad \text{for all large } n.$$

Hence, if A_n adequately characterizes the set of problematic densities, we would expect to find that its prior probability goes to zero, which

indeed turns out to be the case. To see this, observe that $\tilde{q}_n = q$ for all large n, and that for any $p \in \mathcal{M}$ there exists an n_p such that

$$c_1 D_\alpha(p\|q) \geq \frac{L(p) - L(q)}{n} \qquad \text{for all } n \geq n_p.$$

As a consequence, $\pi(A_n) \to 0$. Note however, that the prior probability of A_n may go to zero arbitrarily slowly, so from the consistency conditions we do not get any rate of convergence.

5.7 Discussion

The main result of this chapter, Theorem 5.2, may be considered in its own right, as a convergence result for the MDL estimator. Instead of the light-tails condition, which has previously been suggested, it shows that the standard MDL estimator converges at a rate determined by the index of resolvability if the prior probability of the set A_n is sufficiently small. In the previous section it was also found that the prior probability of A_n goes to zero under conditions under which MDL is known to be consistent.

To study convergence properties of the MDL estimator, the next step would be to investigate whether Theorem 5.2 can be strengthened to convergence in expectation, preferably for finite samples like in Theorem 1.4. However, our motivation in Chapter 1 was much more ambitious: our goal was to gain insight into whether data compression can be made a fundamental notion, which gives a robust interpretation to statistical inference that does not break down when standard, but hard to verify, assumptions fail. What do our findings say about this?

We have seen that the standard MDL estimator may be inconsistent if no conditions are imposed on the density code lengths. This is worrying, because it suggests that good data compression may still lead to bad statistical inference. However, we have already seen in Chapter 2 that standard MDL methods do not necessarily achieve the best compression. We will now present an informal data compression argument which suggests that suboptimal compression may be the problem in the inconsistency examples from this chapter as well.

Let us assume that the MDL estimator is applied to a finite set of densities $\ddot{\mathcal{M}} = \{p_1, \ldots, p_m\}$ which together cover a larger model $\mathcal{M}$, as in Section 1.3.5 of Chapter 1. To achieve the best compression, this

set should be as small as possible, so that each density in $\ddot{\mathcal{M}}$ covers a different neighbourhood of $\mathcal{M}$.

The MDL estimator bases its decision on the two-part code with code lengths

$$L_{\text{2-p}}(x^n) = \min_{p \in \ddot{\mathcal{M}}} L_n(p) - \log p(x^n),$$

which explicitly encodes both the data and the density $\ddot{p}_n \in \ddot{\mathcal{M}}$ that is selected by MDL. However, if $\ddot{p}_n$ provides an accurate summary of x^n, then *all information about* $\ddot{p}_n$ *should also be information about* x^n, and explicitly encoding $\ddot{p}_n$ together with the data should not cost significantly more bits than just encoding the data.

Given the density code lengths used by MDL, a natural way to encode just the data, without encoding $\ddot{p}_n$, is to use the Bayesian universal code with the prior π_n such that $L_n(p) = -\log \pi_n(p)$. The corresponding code length is

$$L_{\text{B}}(x^n) = -\log \sum_{p \in \ddot{\mathcal{M}}} \pi_n(p) p(x^n).$$

As the two-part code explicitly encodes an element from $\ddot{\mathcal{M}}$ and the Bayesian universal code does not, it is not surprising that $L_{\text{2-p}}(x^n) \geq L_{\text{B}}(x^n)$ for all data x^n, as was shown in Chapter 1. But based on the reasoning above, we now also impose the requirement that $L_{\text{2-p}}(x^n)$ should not be *much* larger than $L_{\text{B}}(x^n)$; that is, we require that

$$L_{\text{2-p}}(x^n) \approx L_{\text{B}}(x^n).$$

Rewriting this expression in terms of the corresponding densities as

$$\max_{p \in \ddot{\mathcal{M}}} \frac{\pi_n(p) p(x^n)}{\sum_p \pi_n(p) p(x^n)} \approx 1,$$

we get a different interpretation: it turns out that we can reinterpret it as saying that the Bayesian posterior distribution should converge on a single density in $\ddot{\mathcal{M}}$. This line of reasoning suggests that the MDL estimator achieves suboptimal compression unless the corresponding posterior distribution converges. Indeed, in the inconsistency examples from Section 5.2 it is seen that the posterior does not converge, because the posterior probability of the true density does not go to zero as n

grows. Hence we conjecture that convergence of the Bayesian posterior distribution is necessary for consistency of the MDL estimator.

If this is really the case, then the results of this chapter have an interesting implication: they would show that Condition 5.1 would be sufficient, not just for convergence of the MDL estimator, but also for convergence of the Bayesian posterior distribution. Verifying this would be an interesting direction for future work.

5.8 Future Work

Much is known about convergence of the Bayesian posterior distribution. For example, Barron, Schervish, and Wasserman [1999] and Ghosal, Ghosh and Van der Vaart [2000] prove concentration of the posterior on Hellinger neighbourhoods of the true density. Their results require that the prior puts non-negligible probability mass on a neighbourhood of the true density, and that (except for a set of negligible prior probability) the model can be covered by a sufficiently small number of ε-balls as $\varepsilon \to 0$. The analysis by Barron et al. depends on an application of the Chernoff bound with a fixed choice of $\gamma = 1/2$. One would therefore expect that it would benefit from a varying choice of γ, as in the proof of Lemma 5.2. This might lead to weaker conditions for posterior concentration that only require the prior probability of A_n to be sufficiently small.

Convergence of the posterior has also been studied by Zhang [2006], who obtains rates of convergence under a single requirement, which is essentially the light-tails condition. The light-tails condition (with $\lambda = 2$) is also encountered by Walker [2004], who provides some additional discussion. As we have found that for MDL convergence the light-tails condition could be relaxed to conditions on the prior probability of A_n, these results provide another indication that it might be possible to obtain convergence of the posterior if the prior probability of A_n is sufficiently small.

Zhang further analyses convergence of the MDL estimator. His techniques might form a starting point to strengthen the convergence in probability shown by Theorem 5.2 to convergence in expectation.

Finally, there may be an interesting connection to an inconsistency result for MDL model selection by Csiszár and Shields [2000], who show that using MDL to determine the order of a Markov chain will

select unboundedly large orders if the data are generated uniformly at random, even though the true distribution of the data can be represented as a first order Markov chain. Although Csiszár and Shields only analyse Bayesian and NML universal codes, it seems plausible that their result would extend to two-part universal codes. In that case the results from this chapter apply, and an analysis of the prior probability of A_n might give more insight into why MDL selects overly complex models.

Chapter 6

Rényi Divergence

Rényi divergence is related to Rényi entropy much like information divergence (also called Kullback-Leibler divergence or relative entropy) is related to Shannon's entropy, and comes up in many settings. It was introduced by Rényi as a measure of information that satisfies almost the same axioms as information divergence. We review the most important properties of Rényi divergence. While some of our results are already known for finite spaces or follow easily from existing results about f-divergences, our contribution here is (a) to extend all results to the continuous case; and (b) to provide a unified overview of all relevant properties, with direct proofs that rely only on general results from measure theory.

6.1 Introduction

The Shannon entropy and the information divergence (also known as relative entropy or Kullback-Leibler divergence) are perhaps the two most fundamental quantities in information theory and its applications. Because of their success, there have been many attempts to generalize these concepts, and in the literature one will find numerous entropy and divergence measures. Most of these quantities have never found any applications, and almost none of them have found an interpretation in terms of coding. The most important exceptions are the Rényi entropy and Rényi divergence [Rényi, 1961]: Harremoës [2006] and Grünwald [2007, p. 649] provide an operational characterization of Rényi divergence as the number of bits by which a mixture of two codes can be compressed; and Csiszár [1995] gives an operational characterization of Rényi divergence as the cut-off rate in block coding and

hypothesis testing. Rényi divergence appears as a crucial tool in proofs of convergence of minimum description length and Bayesian estimators, both in parametric and nonparametric models (see the previous chapter and [Zhang, 2006, Haussler and Opper, 1997]). It is also closely related to Hellinger distance, which is commonly used in the analysis of nonparametric density estimation [Le Cam, 1973, Birgé, 1986, Van de Geer, 1993].

Rényi entropy is well studied [Aczél and Daróczy, 1975, Ben-Bassat and Raviv, 1978], but although Rényi divergence appears in many computations, it has hitherto not been studied systematically. This chapter is intended as a reference document, which treats the basic properties of Rényi divergence in detail.

Rényi's Information Measures For finite alphabets, the *Rényi divergence* of positive order $\alpha \neq 1$ of a probability distribution $P = (p_1, \ldots, p_n)$ from another distribution $Q = (q_1, \ldots, q_n)$ is

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \sum_{i=1}^{n} p_i^\alpha q_i^{1-\alpha}, \tag{6.1}$$

where, for $\alpha > 1$, we read $p_i^\alpha q_i^{1-\alpha}$ as $p_i^\alpha / q_i^{(\alpha-1)}$ and adopt the convention that $0/0 = 0$ and $x/0 = \infty$ for $x > 0$. The *Rényi entropy*

$$H_\alpha(P) = \frac{1}{1 - \alpha} \log \sum_{i=1}^{n} p_i^\alpha$$

can be expressed in terms of the Rényi divergence of P from the uniform distribution $U = (1/n, \ldots, 1/n)$:

$$H_\alpha(P) = H_\alpha(U) - D_\alpha(P\|U) = \log n - D_\alpha(P\|U).$$

As α tends to 1, the Rényi entropy tends to the Shannon entropy and the Rényi divergence tends to the information divergence, so we recover a well-known relation.

There is another way of relating Rényi entropy and Rényi divergence, in which entropy is considered as self-information. Let X denote a discrete random variable with distribution P, and let P_{diag} be the distribution of (X, X). Then

$$H_\alpha(P) = D_{2-\alpha}\left(P_{\text{diag}} \| P \times P\right).$$

For α tending to 1 the right-hand side tends to the mutual information between X and itself, and again a well-known formula is recovered.

Special Orders Although one can define the Rényi divergence of any order, certain values have wider application than others. Of particular interest are the values 0, 1/2, 1, 2, and ∞. The values 0, 1, and ∞ are *extended orders* in the sense that Rényi divergence of these orders cannot be calculated by plugging into (6.1). Instead their definitions are determined by continuity in α. This leads to defining Rényi divergence of order 1 as the information divergence. For order 0 it becomes $-\log Q(\{i \mid p_i > 0\})$, which is closely related to absolute continuity and mutual singularity of the distributions P and Q (see Section 6.5.7). And for order ∞, Rényi divergence is defined as $\log \max_i p_i/q_i$, which is related to the *separation distance*, used by Aldous and Diaconis [1987] to bound the rate of convergence to the stationary distribution for certain Markov chains. Only for $\alpha = 1/2$ is Rényi divergence symmetric in its arguments. Although not itself a metric, it is a function of the square of the Hellinger distance $\mathrm{Hel}^2(P,Q) = \sum_{i=1}^n (\sqrt{p_i} - \sqrt{q_i})^2$ [Gibbs and Su, 2002]:

$$D_{1/2}(P\|Q) = -2\log\left(1 - \frac{\mathrm{Hel}^2(P,Q)}{2}\right). \tag{6.2}$$

Similarly, for $\alpha = 2$ it satisfies

$$D_2(P\|Q) = \log\left(1 + \chi^2(P,Q)\right), \tag{6.3}$$

where $\chi^2(P,Q) = \sum_{i=1}^n \frac{(p_i - q_i)^2}{q_i}$ denotes the χ^2-*distance* [Gibbs and Su, 2002]. It will be shown that Rényi divergence is nondecreasing in its order. Therefore, by $\log t \le t - 1$, (6.2) and (6.3) imply that

$$\mathrm{Hel}^2(P,Q) \le D_{1/2}(P\|Q) \le D_1(P\|Q) \le D_2(P\|Q) \le \chi^2(P,Q).$$

Outline The rest of the chapter is organized as follows. In Section 6.2 we extend the definition of Rényi divergence from the formula (6.1) to continuous spaces. One can either define Rényi divergence via an integral or via discretisations. We demonstrate that these definitions are equivalent. In Section 6.3 some basic properties are established, which

are required in the rest of the chapter. For continuous spaces, Rényi divergence extends to the extended orders $0, 1$ and ∞ in the same way as for finite spaces. This is shown in Section 6.4, where we analyse Rényi divergence as a function of its order α. In Section 6.5 we study Rényi divergence as function of the two probability distributions P and Q for fixed α. Finally, Section 6.6 reviews applications of Rényi divergence and provides further references. It includes a connection to hypothesis testing, to which most applications of Rényi divergence are related.

For fixed α, Rényi divergence is related to various forms of *power divergences*, which are in the well-studied class of *f-divergences* [Liese and Vajda, 2006]. Consequently, several of the results we are presenting for fixed α in Sections 6.3 and 6.5 are equivalent to known results about power divergences. To make this presentation self-contained we avoid the use of such connections and only use general results from measure theory.

6.2 Definition of Rényi divergence

Let us introduce the notation used throughout the chapter. We consider (probability) measures on a measurable space $(\mathcal{X}, \mathcal{F})$. Any such measure P is called *absolutely continuous* with respect to another measure Q if $P(A) = 0$ whenever $Q(A) = 0$ for all events $A \in \mathcal{F}$. We will write $P \ll Q$ if P is absolutely continuous with respect to Q and $P \not\ll Q$ otherwise. Alternatively, P and Q may be *mutually singular*, denoted $P \perp Q$, which means that there exists an event $A \in \mathcal{F}$ such that $P(A) = 0$ and $Q(\mathcal{X} \setminus A) = 0$. We will assume that all (probability) measures are absolutely continuous with respect to a common σ-finite measure μ, which is arbitrary in the sense that none of our definitions or results depend on the choice of μ. As we only consider (mixtures of) a countable number of distributions, such a measure μ exists in all cases, so this does not restrict our treatment. For measures denoted by capital letters (e.g. P or Q), we will use the corresponding lower-case letters (e.g. p, q) to refer to their densities with respect to μ. Finally, for any event $A \in \mathcal{F}$, $\mathbf{1}_A$ denotes its indicator function, which is 1 on A and 0 otherwise, and log denotes the natural logarithm.

We will often need to distinguish between the orders for which Rényi divergence can be defined by a generalisation of the formula (6.1) to an integral over densities, and the other orders. This motivates

the following definitions.

Definition 6.1. We call a (finite) real number α a *simple order* if $\alpha > 0$ and $\alpha \neq 1$. The values $0, 1, \infty$ are called *extended orders*.

6.2.1 Definition by Formula

The formula in (6.1), which defines Rényi divergence for simple orders on finite sample spaces, generalises to arbitrary spaces as follows:

Definition 6.2 (Simple Orders)**.** For any simple order α, the *Rényi divergence* of *order* α of a probability distribution P from another distribution Q is defined as

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \int p^\alpha q^{1-\alpha} \, \mathrm{d}\mu, \tag{6.4}$$

where, for $\alpha > 1$, we read $p^\alpha q^{1-\alpha}$ as $p^\alpha / q^{(\alpha-1)}$ and adopt the conventions that $0/0 = 0$ and $x/0 = \infty$ for $x > 0$.

As a consequence of Theorem 6.2 below, this definition does not depend on the choice of μ. In addition, its interpretation of $p^\alpha q^{1-\alpha}$ is such that the *Hellinger integral* $\int p^\alpha q^{1-\alpha} \, \mathrm{d}\mu$ is an f-divergence [Liese and Vajda, 2006], which ensures that the relations to squared Hellinger distance and χ^2-distance from the introduction (Equations 6.2 and 6.3) hold in general, not just for finite sample spaces. For simple orders, we may always change to integration with respect to P:

$$\int p^\alpha q^{1-\alpha} \, \mathrm{d}\mu = \int \left(\frac{q}{p}\right)^{1-\alpha} \mathrm{d}P,$$

and in most cases it is also equivalent to integrate with respect to Q:

$$\int p^\alpha q^{1-\alpha} \, \mathrm{d}\mu = \int \left(\frac{p}{q}\right)^{\alpha} \mathrm{d}Q \qquad (0 < \alpha < 1 \text{ or } P \ll Q).$$

However, if $\alpha > 1$ and $P \not\ll Q$, then $D_\alpha(P\|Q) = \infty$, whereas the integral with respect to Q may be finite.

6.2.2 Definition via Discretisation

For any measure λ on $(\mathcal{X}, \mathcal{F})$, let $\lambda_{|\mathcal{G}}$ denote its restriction to the σ-subalgebra $\mathcal{G} \subseteq \mathcal{F}$. We shall repeatedly use the following result, which is a direct consequence of the Radon-Nikodým theorem [Shiryaev, 1996]:

Proposition 6.1. *Suppose $\lambda \ll \mu$ is a probability distribution, or any countably additive measure such that $\lambda(\mathcal{X}) \leq 1$. Then for any σ-subalgebra $\mathcal{G} \subseteq \mathcal{F}$*

$$\frac{\mathrm{d}\lambda_{|\mathcal{G}}}{\mathrm{d}\mu_{|\mathcal{G}}} = \mathbf{E}\left[\frac{\mathrm{d}\lambda}{\mathrm{d}\mu}\middle|\,\mathcal{G}\right] \qquad (\mu\text{-a.s.})$$

Proof. The function $\mathbf{E}\left[\frac{\mathrm{d}\lambda}{\mathrm{d}\mu}\middle|\,\mathcal{G}\right]$ is $\mathcal{G}$-measurable and satisfies

$$\lambda(A) = \int_A \mathbf{E}\left[\frac{\mathrm{d}\lambda}{\mathrm{d}\mu}\middle|\,\mathcal{G}\right]\mathrm{d}\mu, \qquad A \in \mathcal{G}.$$

As the Radon-Nikodým theorem asserts that these requirements are only satisfied by functions that are equal to $\mathrm{d}\lambda_{|\mathcal{G}}/\mathrm{d}\mu_{|\mathcal{G}}$ except on sets of μ-measure zero, the proposition follows. □

6.2.2.1 Data Processing

Until Section 6.5, let P and Q be fixed distributions on $(\mathcal{X}, \mathcal{F})$. It has been argued that grouping observations together (by considering a coarser σ-algebra), should not increase our ability to distinguish between P and Q under any measure of divergence [Ali and Silvey, 1966]. This is expressed by the *data processing inequality*, which Rényi divergence satisfies:

Theorem 6.1 (Data Processing Inequality)**.** *For any simple order α and any σ-subalgebra $\mathcal{G} \subseteq \mathcal{F}$*

$$D_\alpha(P_{|\mathcal{G}} \| Q_{|\mathcal{G}}) \leq D_\alpha(P\|Q).$$

Proof. Let $\tilde{P}$ denote the absolutely continuous component of P with respect to Q. Then by Proposition 6.1 and Jensen's inequality for con-

ditional expectations

$$\frac{1}{\alpha-1}\log\int\left(\frac{\mathrm{d}\tilde{P}_{|\mathcal{G}}}{\mathrm{d}Q_{|\mathcal{G}}}\right)^{\alpha}\mathrm{d}Q=\frac{1}{\alpha-1}\log\int\left(\mathbf{E}\left[\frac{\mathrm{d}\tilde{P}}{\mathrm{d}Q}\middle|\mathcal{G}\right]\right)^{\alpha}\mathrm{d}Q$$
$$\leq\frac{1}{\alpha-1}\log\int\mathbf{E}\left[\left(\frac{\mathrm{d}\tilde{P}}{\mathrm{d}Q}\right)^{\alpha}\middle|\mathcal{G}\right]\mathrm{d}Q$$
$$=\frac{1}{\alpha-1}\log\int\left(\frac{\mathrm{d}\tilde{P}}{\mathrm{d}Q}\right)^{\alpha}\mathrm{d}Q. \tag{6.5}$$

If $0<\alpha<1$, then $p^{\alpha}q^{1-\alpha}=0$ if $q=0$, so the restriction of P to $\tilde{P}$ does not change the Rényi divergence, and hence the theorem is proved. Alternatively, suppose $\alpha>1$. If $P\ll Q$, then $\tilde{P}=P$ and the theorem again follows from (6.5). If $P\not\ll Q$, then $D_{\alpha}(P\|Q)=\infty$ and the theorem holds as well. □

6.2.2.2 Approximation by Finite Partitions

For any finite or countable partition $\mathcal{P}=\{A_1,A_2,\ldots\}$ of $\mathcal{X}$, let $P_{|\mathcal{P}}\equiv P_{|\sigma(\mathcal{P})}$ and $Q_{|\mathcal{P}}\equiv Q_{|\sigma(\mathcal{P})}$ denote the restrictions of P and Q to the σ-algebra generated by $\mathcal{P}$.

Theorem 6.2. *For any simple order α*

$$D_{\alpha}(P\|Q)=\sup_{\mathcal{P}}D_{\alpha}(P_{|\mathcal{P}}\|Q_{|\mathcal{P}}), \tag{6.6}$$

where the supremum is over all finite partitions $\mathcal{P}\subseteq\mathcal{F}$.

This shows that it would be equivalent to first define Rényi divergence for finite sample spaces and then extend the definition to arbitrary sample spaces using (6.6). As for finite sample spaces Rényi divergence does not depend on the choice of dominating measure μ, Theorem 6.2 implies that it does not depend on the choice of μ in general.

Proof of Theorem 6.2. By the data processing inequality

$$\sup_{\mathcal{P}}D_{\alpha}(P_{|\mathcal{P}}\|Q_{|\mathcal{P}})\leq D_{\alpha}(P\|Q).$$

To show the converse inequality, consider for any $\varepsilon > 0$ a discretisation of the densities p and q into a countable number of bins

$$B^\varepsilon_{m,n} = \{x \in \mathcal{X} \mid e^{m\varepsilon} \le p(x) < e^{(m+1)\varepsilon}, e^{n\varepsilon} \le q(x) < e^{(n+1)\varepsilon}\},$$

where $n, m \in \{-\infty, \dots, -1, 0, 1, \dots\}$. Let $\mathcal{Q}^\varepsilon = \{B^\varepsilon_{m,n}\}$ and $\mathcal{F}^\varepsilon = \sigma(\mathcal{Q}^\varepsilon) \subseteq \mathcal{F}$ be the corresponding partition and σ-algebra, and let $p_\varepsilon = \mathrm{d}P_{|\mathcal{Q}^\varepsilon}/\mathrm{d}\mu$ and $q_\varepsilon = \mathrm{d}Q_{|\mathcal{Q}^\varepsilon}/\mathrm{d}\mu$ be the densities of P and Q restricted to $\mathcal{F}^\varepsilon$. Then by Proposition 6.1

$$\frac{q_\varepsilon}{p_\varepsilon} = \frac{\mathbf{E}[q \mid \mathcal{B}^\varepsilon]}{\mathbf{E}[p \mid \mathcal{B}^\varepsilon]} \le \frac{q}{p} e^{2\varepsilon} \qquad (P\text{-a.s.})$$

It follows that

$$\frac{1}{\alpha - 1} \log \int \left(\frac{q_\varepsilon}{p_\varepsilon}\right)^{1-\alpha} \mathrm{d}P \ge \frac{1}{\alpha - 1} \log \int \left(\frac{q}{p}\right)^{1-\alpha} \mathrm{d}P - 2\varepsilon,$$

and hence the supremum over all countable partitions is large enough:

$$\sup_{\substack{\text{countable } \mathcal{Q} \\ \sigma(\mathcal{Q}) \subseteq \mathcal{F}}} D_\alpha(P_{|\mathcal{Q}} \| Q_{|\mathcal{Q}}) \ge \sup_{\varepsilon > 0} D_\alpha(P_{|\mathcal{Q}^\varepsilon} \| Q_{|\mathcal{Q}^\varepsilon}) \ge D_\alpha(P \| Q).$$

It remains to show that the supremum over finite partitions is at least as large. To this end, suppose $\mathcal{Q} = \{B_1, B_2, \dots\}$ is any countable partition and let $\mathcal{P}_n = \{B_1, \dots, B_{n-1}, \bigcup_{i \ge n} B_i\}$. Then by

$$P\Big(\bigcup_{i \ge n} B_i\Big)^\alpha Q\Big(\bigcup_{i \ge n} B_i\Big)^{1-\alpha} \ge 0 \qquad (\alpha > 1),$$

$$\lim_{n \to \infty} P\Big(\bigcup_{i \ge n} B_i\Big)^\alpha Q\Big(\bigcup_{i \ge n} B_i\Big)^{1-\alpha} = 0 \qquad (0 < \alpha < 1),$$

we find that

$$\begin{aligned} \lim_{n \to \infty} D_\alpha(P_{|\mathcal{P}_n} \| Q_{|\mathcal{P}_n}) &= \lim_{n \to \infty} \frac{1}{\alpha - 1} \log \sum_{B \in \mathcal{P}_n} P(B)^\alpha Q(B)^{1-\alpha} \\ &\ge \lim_{n \to \infty} \frac{1}{\alpha - 1} \log \sum_{i=1}^{n-1} P(B_i)^\alpha Q(B_i)^{1-\alpha} = D_\alpha(P_{|\mathcal{Q}} \| Q_{|\mathcal{Q}}), \end{aligned}$$

which completes the proof. □

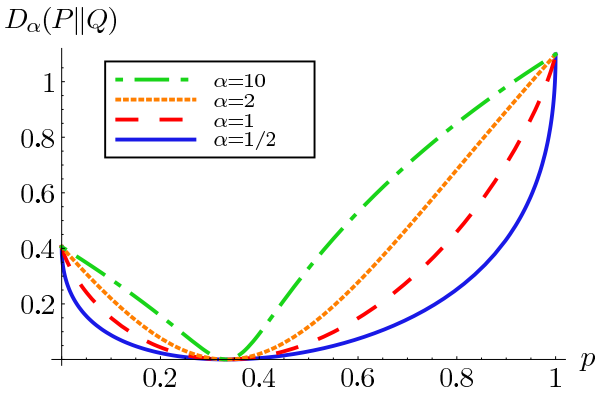


Figure 6.1: Rényi divergence as a function of $P = (p, 1-p)$ for $Q = (1/3, 2/3)$

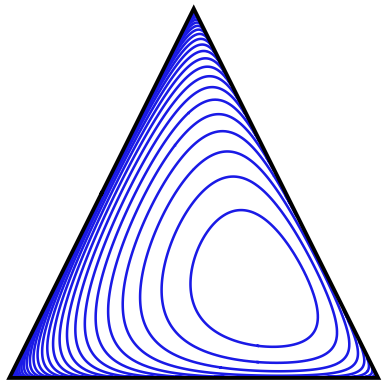


Figure 6.2: Level curves of $D_{1/2}(P\|Q)$ for fixed Q as P ranges over the simplex of distributions on a three-element set

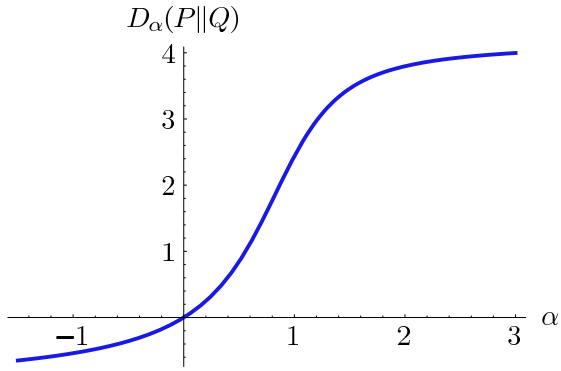


Figure 6.3: Rényi divergence as a function of its order for fixed distributions

6.3 Basic Properties for Simple Orders

Consider Figures 6.1, 6.2 and 6.3. They show $D_\alpha(P\|Q)$ as a function of P for sample spaces containing two or three elements, and as a function of α for fixed P and Q. The figures suggest some basic properties. In particular, for $\alpha > 0$ the plotted divergences are nonnegative and zero only when P equals Q. They are also increasing and continuous in α. Let us verify that, under suitable conditions, these properties always hold for simple orders. Proofs for the extended orders are given later.

Theorem 6.3 (Positivity)**.** *For any simple order α*

$$D_\alpha(P\|Q) \geq 0.$$

Equality holds (i.e. $D_\alpha(P\|Q) = 0$) if and only if $P = Q$.

Proof. By Jensen's inequality

$$\begin{aligned}\frac{1}{\alpha-1}\log\int p^\alpha q^{1-\alpha}\,\mathrm{d}\mu &= \frac{1}{\alpha-1}\log\int\left(\frac{q}{p}\right)^{1-\alpha}\mathrm{d}P\\ &\geq \frac{1-\alpha}{\alpha-1}\log\int\frac{q}{p}\,\mathrm{d}P \geq 0.\end{aligned}$$

Equality holds if and only if q/p is constant P-a.s. (first inequality) and $Q \ll P$ (second inequality), which together is equivalent to $P = Q$. □

Theorem 6.4 (Increasing in the Order)**.** *For simple orders α, the Rényi divergence $D_\alpha(P\|Q)$ is nondecreasing in α. On*

$$A = \{\textit{simple orders } \alpha \mid 0 < \alpha < 1 \textit{ or } D_\alpha(P\|Q) < \infty\}$$

it is constant if and only if q/p is constant P-a.s.

Proof. Let $\alpha < \beta$ be simple orders. Then for $x \geq 0$ the function $x \mapsto x^{(\alpha-1)/(\beta-1)}$ is strictly convex if $\alpha < 1$ and strictly concave if $\alpha > 1$. Therefore by Jensen's inequality

$$\begin{aligned}\frac{1}{\alpha-1}\log\int p^\alpha q^{1-\alpha}\,\mathrm{d}\mu &= \frac{1}{\alpha-1}\log\int\left(\frac{q}{p}\right)^{(1-\beta)\frac{\alpha-1}{\beta-1}}\mathrm{d}P \\ &\leq \frac{1}{\beta-1}\log\int\left(\frac{q}{p}\right)^{(1-\beta)}\mathrm{d}P.\end{aligned}$$

On A, $\int (q/p)^{(1-\beta)}\,\mathrm{d}P$ is finite, so that Jensen's inequality holds with equality if and only if $(q/p)^{1-\beta}$ is constant P-a.s., which is equivalent to the claim of the theorem. □

Theorem 6.5 (Continuous in the Order)**.** *The Rényi divergence $D_\alpha(P\|Q)$ is continuous in α on*

$$A = \{\textit{simple orders } \alpha \mid 0 < \alpha < 1 \textit{ or } D_\alpha(P\|Q) < \infty\}.$$

The theorem follows from the following lemma, applied with $\beta \in A$:

Lemma 6.1. *For any sequence $\alpha_1, \alpha_2, \ldots \in A$ such that $\alpha_n \to \beta \in A \cup \{0, 1\}$*

$$\lim_{n\to\infty}\int p^{\alpha_n}q^{1-\alpha_n}\,\mathrm{d}\mu = \int \lim_{n\to\infty} p^{\alpha_n}q^{1-\alpha_n}\,\mathrm{d}\mu. \tag{6.7}$$

The proof extends a proof by Shiryaev [1996, pp. 366–367].

Proof. We will verify the conditions for the dominated convergence theorem, from which (6.7) follows. First suppose $0 \leq \beta < 1$. Then $0 < \alpha_n < 1$ for all sufficiently large n. In this case $p^{\alpha_n}q^{1-\alpha_n}$, which is never negative, does not exceed $\alpha_n p + (1-\alpha_n)q \leq p + q$, and the dominated convergence theorem applies because $\int (p+q)\,\mathrm{d}\mu = 2 < \infty$. Secondly, suppose $\beta \geq 1$ and assume without loss of generality that

$\alpha_n > 0$. Then there exists a $\gamma \geq \beta$ such that $\gamma \in A \cup \{1\}$ and $\alpha_n \leq \gamma$ for all sufficiently large n. If $\gamma = 1$, then $\alpha_n < 1$ and we are done by the same argument as above. So suppose $\gamma > 1$. Then convexity of $p^{\alpha_n} q^{1-\alpha_n}$ in α_n implies that for $\alpha_n \leq \gamma$

$$p^{\alpha_n} q^{1-\alpha_n} \leq (1 - \frac{\alpha_n}{\gamma}) p^0 q^1 + \frac{\alpha_n}{\gamma} p^\gamma q^{1-\gamma} \leq q + p^\gamma q^{1-\gamma}.$$

Since $\int q \, \mathrm{d}\mu = 1$, it remains to show that $\int p^\gamma q^{1-\gamma} \, \mathrm{d}\mu < \infty$, which is implied by $\gamma > 1$ and $D_\gamma(P\|Q) < \infty$. □

6.4 Extended Orders: Varying the Order

As for finite alphabets, continuity considerations lead to the following extensions of Rényi divergence to orders for which it cannot be defined using (6.4).

Definition 6.3 (Extended Orders)**.** The *Rényi divergences* of orders 0 and 1 are defined as

$$D_0(P\|Q) = \lim_{\alpha \downarrow 0} D_\alpha(P\|Q) = -\log Q(p > 0),$$
$$D_1(P\|Q) = \lim_{\alpha \uparrow 1} D_\alpha(P\|Q) = D(P\|Q),$$

and of order ∞ as

$$D_\infty(P\|Q) = \lim_{\alpha \uparrow \infty} D_\alpha(P\|Q) = \log \sup_{A \in \mathcal{F}} \frac{P(A)}{Q(A)},$$

with the convention that $0/0 = 0$.

Here $D(P\|Q)$ denotes the *information divergence* of a probability distribution P from another distribution Q, which is defined as

$$D(P\|Q) = \int p \log \frac{p}{q} \, \mathrm{d}\mu,$$

with the conventions that $0 \log(0/q) = 0$ and $p \log(p/0) = \infty$ if $p > 0$. Consequently, $D(P\|Q) = \infty$ if $P \not\ll Q$. Our definition of D_0 follows Csiszár [1995]. It differs from Rényi's original definition, which equals (6.4) with $\alpha = 0$ plugged in [Rényi, 1961] and is therefore always zero. As illustrated by Section 6.5.7, the present definition is more interesting.

Let us verify that the properties for simple orders from the previous section also hold for the extended orders, and that the limits in Definition 6.3 equal their corresponding closed forms as claimed. Taking the limits in Definition 6.3 as our basic definitions, we directly find the following:

Theorem 6.6 (Increasing in the Order)**.** *For $0 \leq \alpha \leq \infty$ the Rényi divergence $D_\alpha(P\|Q)$ is nondecreasing in α. On*

$$A = \{0 \leq \alpha \leq \infty \mid 0 \leq \alpha \leq 1 \text{ or } D_\alpha(P\|Q) < \infty\}$$

it is constant if and only if q/p is constant P-a.s.

Proof. From the simple orders (Theorem 6.4), the result extends to the extended orders by the following observations:

$$\begin{aligned} D_0(P\|Q) &= \inf_{0<\alpha<1} D_\alpha(P\|Q), \\ D_1(P\|Q) &= \sup_{0<\alpha<1} D_\alpha(P\|Q) \leq \inf_{\alpha>1} D_\alpha(P\|Q), \\ D_\infty(P\|Q) &= \sup_{\alpha>1} D_\alpha(P\|Q). \end{aligned}$$

□

And the limits equal their corresponding closed form expressions:

Theorem 6.7 ($\alpha = 0$)**.**

$$\lim_{\alpha\downarrow 0} D_\alpha(P\|Q) = -\log Q(p > 0).$$

Proof. By Lemma 6.1 and the fact that $\lim_{\alpha\downarrow 0} p^\alpha q^{1-\alpha} = \mathbf{1}_{\{p>0\}} q$. □

Theorem 6.8 ($\alpha = 1$)**.**

$$\lim_{\alpha\uparrow 1} D_\alpha(P\|Q) = D(P\|Q). \tag{6.8}$$

Moreover, if $D(P\|Q) = \infty$ or there exists a $\beta > 1$ such that $D_\beta(P\|Q) < \infty$, then also

$$\lim_{\alpha\downarrow 1} D_\alpha(P\|Q) = D(P\|Q). \tag{6.9}$$

It is possible however that $D_\alpha(P\|Q) = \infty$ for all $\alpha > 1$, but $D(P\|Q) < \infty$, such that (6.9) does not hold. This situation occurs, for example, if P is doubly exponential on $\mathcal{X} = \mathbb{R}$ with density $p(x) = e^{-2|x|}$ and Q

is standard normal with density $q(x) = e^{-x^2/2}/\sqrt{2\pi}$. (Liese and Vajda [2006] have previously used these distributions in a similar example.) In this case there is no way to make Rényi divergence continuous in α at $\alpha = 1$, and we opt to define D_1 as the limit from below, such that it always equals the information divergence.

The proof of Theorem 6.8 requires an intermediate lemma:

Lemma 6.2. *For any $x > 1/2$*

$$(x-1)\left(1+\frac{1-x}{2}\right) \le \log x \le x - 1.$$

Proof. By Taylor's theorem with Cauchy's remainder term we have for any positive x that $\log x = x - 1 - \frac{(x-e)(x-1)}{2e^2} = (x-1)(1+\frac{e-x}{2e^2})$ for some e between x and 1. As $\frac{e-x}{2e^2}$ is increasing in e for $x > 1/2$, the lemma follows. □

Proof of Theorem 6.8. Suppose $P \not\ll Q$. Then

$$D(P\|Q) = D_\beta(P\|Q) = \infty$$

for all $\beta > 1$, so (6.9) holds. And (6.8) follows by

$$\begin{aligned}\lim_{\alpha\uparrow 1}\frac{1}{\alpha-1}\log\int p^\alpha q^{1-\alpha}\,\mathrm{d}\mu &\ge \lim_{\alpha\uparrow 1}\frac{1}{\alpha-1}\log\int (\mathbf{1}_{\{q>0\}}p)^\alpha\,\mathrm{d}\mu \\ &\ge \lim_{\alpha\uparrow 1}\frac{\alpha}{\alpha-1}\log P(q>0) = \infty = D(P\|Q),\end{aligned}$$

where the second inequality is Jensen's.

Alternatively, suppose $P \ll Q$ and let $x_\alpha = \int p^\alpha q^{1-\alpha}\,\mathrm{d}\mu$. Then $\lim_{\alpha\uparrow 1} x_\alpha = 1$ by Lemma 6.1. Therefore Lemma 6.2 implies that

$$\begin{aligned}\lim_{\alpha\uparrow 1} D_\alpha(P\|Q) &= \lim_{\alpha\uparrow 1}\frac{1}{\alpha-1}\log x_\alpha \\ &= \lim_{\alpha\uparrow 1}\frac{x_\alpha - 1}{\alpha-1} = \lim_{\alpha\uparrow 1}\int_{p,q>0}\frac{p - p^\alpha q^{1-\alpha}}{1-\alpha}\,\mathrm{d}\mu, \qquad (6.10)\end{aligned}$$

where the restriction of the domain of integration is allowed because $q = 0$ implies $p = 0$ (μ-a.s.) by $P \ll Q$. Convexity of $p^\alpha q^{1-\alpha}$ in α implies that its derivative, $p^\alpha q^{1-\alpha}\log\frac{p}{q}$, is nondecreasing and therefore for $p, q > 0$

$$\frac{p - p^\alpha q^{1-\alpha}}{1-\alpha} = \frac{1}{1-\alpha}\int_\alpha^1 p^z q^{1-z}\log\frac{p}{q}\,\mathrm{d}z$$

is nondecreasing in α, and

$$\frac{p - p^\alpha q^{1-\alpha}}{1-\alpha} \geq \frac{p - p^0 q^{1-0}}{1-0} = p - q.$$

As $\int_{p,q>0}(p-q)\,\mathrm{d}\mu > -\infty$, it follows by the monotone convergence theorem that

$$\begin{aligned}\lim_{\alpha\uparrow 1}\int_{p,q>0}\frac{p - p^\alpha q^{1-\alpha}}{1-\alpha}\,\mathrm{d}\mu &= \int_{p,q>0}\lim_{\alpha\uparrow 1}\frac{p - p^\alpha q^{1-\alpha}}{1-\alpha}\,\mathrm{d}\mu \\ &= \int_{p,q>0} p\log\frac{p}{q}\,\mathrm{d}\mu = D(P\|Q),\end{aligned}$$

which together with (6.10) proves (6.8).

If $D(P\|Q) = \infty$, then $D_\beta(P\|Q) \geq D(P\|Q) = \infty$ for all $\beta > 1$ and (6.9) holds. It remains to prove (6.9) if there exists a $\beta > 1$ such that $D_\beta(P\|Q) < \infty$. In this case, arguments similar to the ones above imply that

$$\lim_{\alpha\downarrow 1} D_\alpha(P\|Q) = \lim_{\alpha\downarrow 1}\int_{p,q>0}\frac{p^\alpha q^{1-\alpha} - p}{\alpha - 1}\,\mathrm{d}\mu \tag{6.11}$$

and $\frac{p^\alpha q^{1-\alpha}-p}{\alpha-1}$ is increasing in α. Therefore

$$\frac{p^\alpha q^{1-\alpha} - p}{\alpha - 1} \leq \frac{p^\beta q^{1-\beta} - p}{\beta - 1} \leq \frac{p^\beta q^{1-\beta}}{\beta - 1}$$

and, as $\int_{p,q>0}\frac{p^\beta q^{1-\beta}}{\beta-1}\,\mathrm{d}\mu < \infty$ is implied by $D_\beta(P\|Q) < \infty$, it follows by the monotone convergence theorem that

$$\begin{aligned}\lim_{\alpha\downarrow 1}\int_{p,q>0}\frac{p^\alpha q^{1-\alpha} - p}{\alpha - 1}\,\mathrm{d}\mu &= \int_{p,q>0}\lim_{\alpha\downarrow 1}\frac{p^\alpha q^{1-\alpha} - p}{\alpha - 1}\,\mathrm{d}\mu \\ &= \int_{p,q>0} p\log\frac{p}{q}\,\mathrm{d}\mu = D(P\|Q),\end{aligned}$$

which together with (6.11) completes the proof. □

Theorem 6.9 ($\alpha = \infty$)**.**

$$\lim_{\alpha\uparrow\infty} D_\alpha(P\|Q) = \log\sup_{A\in\mathcal{F}}\frac{P(A)}{Q(A)},$$

with the convention that $0/0 = 0$.

Proof. For finite $\mathcal{X}$

$$\begin{aligned} D_\infty(P\|Q) &= \lim_{\alpha\uparrow\infty} \frac{1}{\alpha-1} \log \sum_{i=1}^{|\mathcal{X}|} p_i^\alpha q_i^{1-\alpha} \\ &= \log \max_i \frac{p_i}{q_i} = \log \max_{A\subseteq\mathcal{X}} \frac{P(A)}{Q(A)}. \end{aligned}$$

This extends to arbitrary spaces by Theorem 6.2:

$$\begin{aligned} D_\infty(P\|Q) &= \sup_{\alpha<\infty} \sup_{\mathcal{P}} D_\alpha(P_{|\mathcal{P}}\|Q_{|\mathcal{P}}) = \sup_{\mathcal{P}} \sup_{\alpha<\infty} D_\alpha(P_{|\mathcal{P}}\|Q_{|\mathcal{P}}) \\ &= \sup_{\mathcal{P}} \log \max_{A\in\mathcal{P}} \frac{P(A)}{Q(A)} = \log \sup_{A\in\mathcal{F}} \frac{P(A)}{Q(A)}, \end{aligned}$$

where $\mathcal{P}$ ranges over all finite partitions in $\mathcal{F}$. □

Theorem 6.10 (Continuous in the Order)**.** *The Rényi divergence $D_\alpha(P\|Q)$ is continuous in α on*

$$A = \{0 \le \alpha \le \infty \mid 0 \le \alpha \le 1 \text{ or } D_\alpha(P\|Q) < \infty\}.$$

Proof. Theorems 6.7, 6.8 and 6.9 extend Theorem 6.5 to the extended orders. □

6.5 Extended Orders: Fixed Order

In this section we fix the order α and study properties of Rényi divergence as P and Q are varied. We first extend the data processing inequality and nonnegativity to the extended orders, and then consider convexity and continuity properties.

6.5.1 Data Processing and Positivity

Theorem 6.11 (Data Processing Inequality)**.** *For any order $0 \le \alpha \le \infty$ and any σ-subalgebra $\mathcal{G} \subseteq \mathcal{F}$*

$$D_\alpha(P_{|\mathcal{G}}\|Q_{|\mathcal{G}}) \le D_\alpha(P\|Q). \tag{6.12}$$

Proof. By Theorem 6.1, (6.12) holds for the simple orders. Let β be any extended order and let $\alpha_n \to \beta$ be an arbitrary sequence of simple

orders that converges to β, from above if $\beta = 0$ and from below if $\beta \in \{1, \infty\}$. Then

$$D_\beta(P_{|\mathcal{G}} \| Q_{|\mathcal{G}}) = \lim_{n\to\infty} D_{\alpha_n}(P_{|\mathcal{G}} \| Q_{|\mathcal{G}}) \le \lim_{n\to\infty} D_{\alpha_n}(P\|Q) = D_\beta(P\|Q). \quad \square$$

Theorem 6.12 (Positivity)**.** *For any order* $0 \le \alpha \le \infty$

$$D_\alpha(P\|Q) \ge 0.$$

For $\alpha > 0$, $D_\alpha(P\|Q) = 0$ *if and only if* $P = Q$. *For* $\alpha = 0$, $D_\alpha(P\|Q) = 0$ *if and only if* $Q \ll P$.

Proof. Theorem 6.3 shows that the theorem holds for all simple orders. This extends to $\beta \in \{1, \infty\}$ by $D_\beta(P\|Q) = \sup_{\alpha<\beta} D_\alpha(P\|Q)$. For $\alpha = 0$ it can be verified directly that $-\log Q(p > 0) \ge 0$, with equality if and only if $Q \ll P$. $\square$

6.5.2 Convexity

Figures 6.1 and 6.2 suggest that Rényi divergence is convex in its first argument for small α, but not for large α. This is in agreement with the well-known fact that it is jointly convex in the pair (P, Q) for $\alpha = 1$. It turns out that joint convexity extends to $\alpha < 1$, but not to $\alpha > 1$, as noted by Csiszár [1995]. Our proof generalises the proof for $\alpha = 1$ by Cover and Thomas [1991].

Theorem 6.13. *For any order* $0 \le \alpha \le 1$ *Rényi divergence is jointly convex in its arguments. That is, for any two pairs of probability distributions* (P_0, Q_0) *and* (P_1, Q_1)*, and any* $0 < \lambda < 1$

$$D_\alpha\Big((1-\lambda)P_0 + \lambda P_1 \| (1-\lambda)Q_0 + \lambda Q_1\Big) \le (1-\lambda)D_\alpha(P_0\|Q_0) + \lambda D_\alpha(P_1\|Q_1). \quad (6.13)$$

Equality holds if and only if

$\alpha = 0$*:* $D_0(P_0\|Q_0) = D_0(P_1\|Q_1)$,
$p_0 = 0 \Rightarrow p_1 = 0$ *(Q_0-a.s.) and* $p_1 = 0 \Rightarrow p_0 = 0$ *(Q_1-a.s.);*
$0 < \alpha < 1$*:* $D_\alpha(P_0\|Q_0) = D_\alpha(P_1\|Q_1)$ *and* $p_0 q_1 = p_1 q_0$ *(μ-a.s.);*
$\alpha = 1$*:* $p_0 q_1 = p_1 q_0$ *(μ-a.s.)*

Proof. Let $P_\lambda = (1-\lambda)P_0 + \lambda P_1$ and $Q_\lambda = (1-\lambda)Q_0 + \lambda Q_1$ and first suppose that $\alpha = 0$. Then

$$\begin{aligned}(1-\lambda)\log Q_0(p_0 > 0) + \lambda \log Q_1(p_1 > 0) \\ \leq \log\left((1-\lambda)Q_0(p_0 > 0) + \lambda Q_1(p_1 > 0)\right) \\ \leq \log Q_\lambda(p_0 > 0 \text{ or } p_1 > 0) = \log Q_\lambda(p_\lambda > 0).\end{aligned}$$

Equality holds if and only if, for the first inequality, $Q_0(p_0 > 0) = Q_1(p_1 > 0)$ and, for the second inequality, $p_1 > 0 \Rightarrow p_0 > 0$ (Q_0-a.s.) and $p_0 > 0 \Rightarrow p_1 > 0$ (Q_1-a.s.) These conditions are equivalent to the equality conditions of the theorem.

Alternatively, suppose $\alpha > 0$. We will show that pointwise

$$\begin{aligned}(1-\lambda)p_0^\alpha q_0^{1-\alpha} + \lambda p_1^\alpha q_1^{1-\alpha} &\leq p_\lambda^\alpha q_\lambda^{1-\alpha} \qquad (0 < \alpha < 1); \\ (1-\lambda)p_0 \log\frac{p_0}{q_0} + \lambda p_1 \log\frac{p_1}{q_1} &\geq p_\lambda \log\frac{p_\lambda}{q_\lambda} \qquad (\alpha = 1),\end{aligned} \tag{6.14}$$

where $p_\lambda = (1-\lambda)p_0 + \lambda p_1$ and $q_\lambda = (1-\lambda)q_0 + \lambda q_1$. For $\alpha = 1$ (6.13) then follows directly; for $0 < \alpha < 1$ (6.13) follows from (6.14) by Jensen's inequality:

$$\begin{aligned}(1-\lambda)\log\int p_0^\alpha q_0^{1-\alpha}\,\mathrm{d}\mu + \lambda\log\int p_1^\alpha q_1^{1-\alpha}\,\mathrm{d}\mu \\ \leq \log\left((1-\lambda)\int p_0^\alpha q_0^{1-\alpha}\,\mathrm{d}\mu + \lambda\int p_1^\alpha q_1^{1-\alpha}\,\mathrm{d}\mu\right).\end{aligned} \tag{6.15}$$

If one of p_0, p_1, q_0 and q_1 is zero, then (6.14) can be verified directly. So assume that they are all positive. Then for $0 < \alpha < 1$ let $f(x) = -x^\alpha$ and for $\alpha = 1$ let $f(x) = x\log x$, such that (6.14) can be written as

$$\frac{(1-\lambda)q_0}{q_\lambda} f\left(\frac{p_0}{q_0}\right) + \frac{\lambda q_1}{q_\lambda} f\left(\frac{p_1}{q_1}\right) \geq f\left(\frac{p_\lambda}{q_\lambda}\right).$$

Equation 6.14 is established by recognising this as an application of Jensen's inequality to the strictly convex function f.

Regardless of whether any of p_0, p_1, q_0 and q_1 is zero, equality holds in (6.14) if and only if $p_0 q_1 = p_1 q_0$. Equality holds in (6.15) if and only if $\int p_0^\alpha q_0^{1-\alpha}\,\mathrm{d}\mu = \int p_1^\alpha q_1^{1-\alpha}\,\mathrm{d}\mu$, which is equivalent to $D_\alpha(P_0\|Q_0) = D_\alpha(P_1\|Q_1)$. □

Joint convexity in P and Q breaks down for $\alpha > 1$. To construct a counterexample, let $P_0 = (p_0, 1-p_0)$, $P_1 = (p_1, 1-p_1)$ and $Q_0 = Q_1 = (q, 1-q)$, with $0 < p_0 < p_1 < 1$. Then, for any $\alpha > 1$, (6.13) is violated for all sufficiently small $q > 0$. Instead of joint convexity in both arguments, however, convexity in the second argument does hold for all α [Csiszár, 1995]:

Theorem 6.14. *For any order $0 \le \alpha \le \infty$ Rényi divergence is convex in its second argument. That is, for any probability distributions P, Q_0 and Q_1*

$$D_\alpha(P\|(1-\lambda)Q_0 + \lambda Q_1) \le (1-\lambda)D_\alpha(P\|Q_0) + \lambda D_\alpha(P\|Q_1) \tag{6.16}$$

for any $0 < \lambda < 1$. For finite α, equality holds if and only if

$$\alpha = 0\text{: } D_0(P_0\|Q_0) = D_0(P_1\|Q_1);$$
$$0 < \alpha < \infty\text{: } q_0 = q_1 \text{ (P-a.s.)}$$

Proof. For $0 \le \alpha \le 1$ this follows from the previous theorem. (For $P_0 = P_1$ the equality conditions reduce to the ones given here.)

For $1 < \alpha < \infty$, let $Q_\lambda = (1-\lambda)Q_0 + \lambda Q_1$ and define $f(x, Q_\lambda) = (p(x)/q_\lambda(x))^{\alpha-1}$. It is sufficient to show that

$$\log \mathbf{E}_{X\sim P}[f(X, Q_\lambda)] \le (1-\lambda)\log \mathbf{E}_{X\sim P}[f(X, Q_0)] + \lambda \log \mathbf{E}_{X\sim P}[f(X, Q_1)].$$

Noting that, for every $x \in \mathcal{X}$, $f(x, Q)$ is log-convex in Q, this is a consequence of the general fact that an expectation over log-convex functions is itself log-convex, which can be shown using Hölder's inequality:

$$\begin{aligned}\mathbf{E}_P[f(X, Q_\lambda)] &\le \mathbf{E}_P[f(X, Q_0)^{1-\lambda} f(X, Q_1)^\lambda] \\ &\le \mathbf{E}_P[f(X, Q_0)]^{1-\lambda}\, \mathbf{E}_P[f(X, Q_1)]^\lambda.\end{aligned}$$

Taking logarithms completes the proof of (6.16). Equality holds in the first inequality if and only if $q_0 = q_1$ (P-a.s.), which is also sufficient for equality in the second inequality. Finally, (6.16) extends to $\alpha = \infty$ by letting α tend to ∞. □

6.5.3 No Pythagorean Inequality

An important result in statistical applications of information theory is the Pythagorean inequality for information divergence (see [Cover and Thomas, 1991, Csiszár, 1975, Topsøe, 2007]). It states that, if $\mathcal{P}$ is a convex set of distributions, Q is any distribution not in $\mathcal{P}$, and $D_{\min} = \inf_{P\in\mathcal{P}} D(P\|Q)$, then there exists a distribution P^* such that

$$D(P\|Q) \geq D(P\|P^*) + D_{\min} \qquad \text{for all } P \in \mathcal{P}.$$

The main use of the Pythagorean inequality lies in its implication that if $P_1, P_2, \ldots$ is a sequence of distributions in $\mathcal{P}$ such that $D(P_n\|Q) \to D_{\min}$, then P_n converges to P^* in the strong sense that $D(P_n\|P^*) \to 0$.

Unfortunately, for $\alpha \neq 1$ Rényi divergence does not satisfy the Pythagorean inequality, as demonstrated by the counterexamples below. We should point to results by Sundaresan [2002], however, who argues that, under regularity conditions, for finite sample spaces a generalisation of Rényi divergence (see [Sundaresan, 2006]) does satisfy a modified Pythagorean inequality, in which every distribution $R \in \{P, Q\}$ is replaced by its *tilted* counterpart

$$R'(x) = \frac{R(x)^\alpha}{\sum_y R(y)^\alpha}.$$

To construct the counterexamples for the ordinary Pythagorean inequality, first consider $0 \leq \alpha < 1$. Let $Q = (1/3, 1/3, 1/3)$ be uniform on three points and let $\mathcal{P} = \{(p_1, p_2, p_3) \mid p_1 = 1/4\}$ be the convex set of distributions with first component fixed at $1/4$. Then $\inf_{P\in\mathcal{P}} D_\alpha(P\|Q)$ is achieved by $P^* = (1/4, 3/8, 3/8)$ and the Pythagorean inequality

$$D_\alpha(P\|Q) \geq D_\alpha(P\|P^*) + D_\alpha(P^*\|Q) \tag{6.17}$$

is violated for $P = (1/4, 0, 3/4)$: if $\alpha > 0$, then (6.17) is equivalent to

$$1 + 3^\alpha \leq \left(\frac{1}{4} + \frac{3}{8}2^\alpha\right)\left(1 + 2\left(\frac{3}{2}\right)^\alpha\right)$$

$$(1 - 2^{1-\alpha})(23^\alpha - 32^\alpha) \leq 0,$$

which is false. If $\alpha = 0$, then $D_\alpha(P\|Q) = -\log 2/3, D_\alpha(P\|P^*) = -\log 5/8$ and $D_\alpha(P^*\|Q) = 0$, and the inequality does not hold either.

Secondly, for $1 < \alpha \leq \infty$ take $Q = (1/3, 1/3, 1/3)$ and $\mathcal{P} = \{(p_1, p_2, p_3) \mid p_1 = 2/3\}$. Then $\inf_{P \in \mathcal{P}} D_\alpha(P\|Q)$ is achieved by $P^* = (2/3, 1/6, 1/6)$ and the Pythagorean inequality is violated for $P = (2/3, 0, 1/3)$: if $\alpha < \infty$, then (6.17) is equivalent to

$$\begin{aligned} 6(1 + 2^\alpha) &\geq (4 + 2^\alpha)(2^{1-\alpha} + 2^\alpha) \\ (2^\alpha - 2)(4^\alpha - 4) &\leq 0, \end{aligned}$$

which is false. If $\alpha = \infty$, then $D_\alpha(P\|Q) = D_\alpha(P\|P^*) = D_\alpha(P^*\|Q) = \log 2$ and the inequality does not hold either.

6.5.4 Continuity

In this section we study continuity properties of the Rényi divergence $D_\alpha(P\|Q)$ of different orders in the pair of probability distributions (P, Q). It turns out that continuity depends on the order α and the topology on the set of all probability distributions.

If the set of probability distributions on $(\mathcal{X}, \mathcal{F})$ is equipped with the *τ-topology*, then convergence of a sequence of probability distributions $P_1, P_2, \ldots$ to a probability distribution Q means that $P_n(A) \to Q(A)$ for any $A \in \mathcal{F}$. Alternatively, one might consider the topology defined by the *total variation distance*

$$V(P, Q) = 2 \sup_{A \in \mathcal{F}} |P(A) - Q(A)| = \int |p - q| \, \mathrm{d}\mu,$$

in which $P_n \to Q$ means that $V(P_n, Q) \to 0$. The total variation topology is stronger than the τ-topology in the sense that convergence in total variation distance implies convergence on any $A \in \mathcal{F}$. The two topologies coincide if the sample space $\mathcal{X}$ is countable. If $\mathcal{X}$ is a metric or topological space one may also consider the *weak topology*, which is a weaker topology than τ, but this topology will not be discussed here.

In general, Rényi divergence is lower semi-continuous for positive orders:

Theorem 6.15. *For any order $0 < \alpha \leq \infty$, $D_\alpha(P\|Q)$ is a lower semi-continuous function of the pair (P, Q) in the τ-topology.*

Proof. Suppose $\mathcal{X} = \{a_1, \ldots, a_k\}$ is finite. Then for any simple order α

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \sum_{i=1}^{k} p_i^\alpha q_i^{1-\alpha},$$

where $p_i = P(a_i)$ and $q_i = Q(a_i)$. If $0 < \alpha < 1$, then $p_i^\alpha q_i^{1-\alpha}$ is continuous in (P, Q). For $1 < \alpha < \infty$, it is only discontinuous at $p_i = q_i = 0$, but there $p_i^\alpha q_i^{1-\alpha} = 0 = \min_{(P,Q)} p_i^\alpha q_i^{1-\alpha}$, so then $p_i^\alpha q_i^{1-\alpha}$ is still lower semi-continuous. These properties carry over to $\sum_{i=1}^k p_i^\alpha q_i^{1-\alpha}$ and thus $D_\alpha(P\|Q)$ is continuous for $0 < \alpha < 1$ and lower semi-continuous for $\alpha > 1$.

A supremum over nonnegative (lower semi-)continuous functions is itself lower semi-continuous. Therefore, for simple orders α, Theorem 6.2 implies that $D_\alpha(P\|Q)$ is lower semi-continuous for arbitrary $\mathcal{X}$. This property extends to the extended orders 1 and ∞ by $D_\beta(P\|Q) = \sup_{\alpha<\beta} D_\alpha(P\|Q)$ for $\beta \in \{1, \infty\}$. □

Moreover, if $0 < \alpha < 1$ and the stronger of the two topologies is assumed, then Rényi divergence is uniformly continuous (which implies that it is continuous).

Theorem 6.16. *For $0 < \alpha < 1$, the Rényi divergence $D_\alpha(P\|Q)$ is a uniformly continuous function of (P, Q) in the total variation topology.*

Lemma 6.3. *Let $0 < \alpha < 1$. Then for all $x, y \geq 0$ and $\varepsilon > 0$*

$$|x^\alpha - y^\alpha| \leq \varepsilon^\alpha + \varepsilon^{\alpha-1}|x - y|.$$

Proof. If $x, y \leq \varepsilon$ or $x = y$ the inequality $|x^\alpha - y^\alpha| \leq \varepsilon^\alpha$ is obvious. So assume that $x > y$ and $x \geq \varepsilon$. Then

$$\frac{|x^\alpha - y^\alpha|}{|x - y|} \leq \frac{|x^\alpha - 0^\alpha|}{|x - 0|} = x^{\alpha-1} \leq \varepsilon^{\alpha-1}.$$ □

Proof of Theorem 6.16. First note that Rényi divergence is a function of the *power divergence* $d_\alpha(P, Q) = \int \left(1 - \left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^\alpha\right) \mathrm{d}Q$:

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log\left(1 - d_\alpha(P, Q)\right).$$

Since $x \mapsto \frac{1}{\alpha-1}\log(1 - x)$ is uniformly continuous, it is sufficient to prove that $d_\alpha(P, Q)$ is a uniformly continuous function of (P, Q).

For any $\varepsilon > 0$ and distributions P_1, P_2 and Q, Lemma 6.3 implies that

$$\begin{aligned}\left|d_\alpha(P_1,Q) - d_\alpha(P_2,Q)\right| &\le \int \left|\left(\frac{\mathrm{d}P_1}{\mathrm{d}Q}\right)^\alpha - \left(\frac{\mathrm{d}P_2}{\mathrm{d}Q}\right)^\alpha\right| \mathrm{d}Q \\ &\le \int \left(\varepsilon^\alpha + \varepsilon^{\alpha-1}\left|\frac{\mathrm{d}P_1}{\mathrm{d}Q} - \frac{\mathrm{d}P_2}{\mathrm{d}Q}\right|\right) \mathrm{d}Q \\ &= \varepsilon^\alpha + \varepsilon^{\alpha-1}\int \left|\frac{\mathrm{d}P_1}{\mathrm{d}Q} - \frac{\mathrm{d}P_2}{\mathrm{d}Q}\right| \mathrm{d}Q \\ &= \varepsilon^\alpha + \varepsilon^{\alpha-1} V(P_1,P_2).\end{aligned}$$

As $d_\alpha(P,Q) = d_{1-\alpha}(Q,P)$, it also follows that

$$\left|d_\alpha(P,Q_1) - d_\alpha(P,Q_2)\right| \le \varepsilon^{1-\alpha} + \varepsilon^{-\alpha} V(Q_1,Q_2)$$

for any Q_1, Q_2 and P. Therefore

$$\begin{aligned}&\left|d_\alpha(P_1,Q_1) - d_\alpha(P_2,Q_2)\right| \\ &\qquad \le \left|d_\alpha(P_1,Q_1) - d_\alpha(P_2,Q_1)\right| + \left|d_\alpha(P_2,Q_1) - d_\alpha(P_2,Q_2)\right| \\ &\qquad \le \varepsilon^\alpha + \varepsilon^{\alpha-1} V(P_1,P_2) + \varepsilon^{1-\alpha} + \varepsilon^{-\alpha} V(Q_1,Q_2),\end{aligned}$$

from which the theorem follows. □

In general the Rényi divergence of order $0 < \alpha < 1$ is not continuous in the τ-topology. To construct a counterexample, let P_n denote the probability distribution on $[0, 2\pi]$ with density $\frac{1+\sin(nx)}{2\pi}$ and let Q_n denote the probability distribution on $[0, 2\pi]$ with density $\frac{1-\sin(nx)}{2\pi}$ for $n = 1, 2, \ldots$ Then $D_\alpha\left(P_n \| Q_n\right)$ does not depend on n, and both P_n and Q_n converge to the uniform distribution U on $[0, 2\pi]$ in the τ-topology. Consequently, $\lim_{n\to\infty} D_\alpha\left(P_n \| Q_n\right) \neq 0 = D_\alpha\left(U \| U\right)$, so in general D_α is not continuous in the τ-topology.

It remains to consider $\alpha = 0$. In this case:

Corollary 6.1. *The Rényi divergence $D_0(P\|Q)$ is an upper semi-continuous function of (P, Q) in the total variation topology.*

Proof. This follows from Theorem 6.16 because $D_0(P\|Q)$ is the infimum of the continuous functions $(P, Q) \mapsto D_\alpha(P\|Q)$ for $0 < \alpha < 1$. □

Finally, the topologies induced by Rényi divergence of any order $0 < \alpha < 1$ are equivalent:

Theorem 6.17. *For any* $0 < \alpha < 1$

$$bD_{1/2}(P\|Q) \leq D_\alpha(P\|Q) \leq cD_{1/2}(P\|Q),$$

where $b = \min\{\alpha/(1-\alpha), 1\}$ *and* $c = \max\{\alpha/(1-\alpha), 1\}$.

This follows from the following symmetry-like property, which may be verified directly.

Proposition 6.2 (Skew Symmetry)**.** *For any* $0 < \alpha < 1$

$$D_\alpha(P\|Q) = \frac{\alpha}{1-\alpha} D_{1-\alpha}(Q\|P).$$

Note that, in particular, Rényi divergence is symmetric for $\alpha = 1/2$, but that skew symmetry does not hold for $\alpha = 0$ and $\alpha = 1$.

Proof of Theorem 6.17. Suppose $\alpha \leq 1/2$. Then skew symmetry, together with monotonicity in α, implies that

$$\begin{aligned} D_{1/2}(P\|Q) \geq D_\alpha(P\|Q) &= \frac{\alpha}{1-\alpha} D_{1-\alpha}(Q\|P) \\ &\geq \frac{\alpha}{1-\alpha} D_{1/2}(Q\|P) = \frac{\alpha}{1-\alpha} D_{1/2}(P\|Q). \end{aligned}$$

Similarly for $\alpha \geq 1/2$

$$D_{1/2}(P\|Q) \leq D_\alpha(P\|Q) \leq \frac{\alpha}{1-\alpha} D_{1/2}(P\|Q).$$

Together these two cases prove the theorem. □

6.5.5 Limit of σ-Algebras

Let P and Q be distributions on $(\mathcal{X}, \mathcal{F})$. As shown by Theorem 6.2, there exists a sequence of finite partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots$ such that

$$D_\alpha(P_{|\mathcal{P}_n}\|Q_{|\mathcal{P}_n}) \uparrow D_\alpha(P\|Q). \tag{6.18}$$

Theorem 6.18 below elaborates on this result. It implies that (6.18) holds for any increasing sequence of partitions $\mathcal{P}_1 \subseteq \mathcal{P}_2 \subseteq \cdots$ that generate σ-algebras converging to $\mathcal{F}$, in the sense that $\mathcal{F} = \sigma\left(\bigcup_{n=1}^{\infty} \mathcal{P}_n\right)$. A corresponding result holds for infinite sequences of increasingly coarse partitions, as shown by Theorem 6.19.

Theorem 6.18 (Increasing). *Let $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}$ be a nondecreasing family of σ-algebras, and let $\mathcal{F}_\infty = \sigma\left(\bigcup_{n=1}^{\infty} \mathcal{F}_n\right)$ be the smallest σ-algebra containing them. Then for any order $0 < \alpha \leq \infty$*

$$\lim_{n\to\infty} D_\alpha(P_{|\mathcal{F}_n} \| Q_{|\mathcal{F}_n}) = D_\alpha(P_{|\mathcal{F}_\infty} \| Q_{|\mathcal{F}_\infty}). \tag{6.19}$$

For $\alpha = 0$, (6.19) does not hold. A counterexample is given after Example 6.1 below.

Lemma 6.4. *Let $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}$ be a nondecreasing family of σ-algebras, and let P and μ be probability distributions on $(\mathcal{X}, \mathcal{F})$ such that $P \ll \mu$. Let p be the density of P with respect to μ. Then the family of random variables $\{X_n\}_{n\geq 1}$ with members $X_n = \mathbf{E}\left[p \middle| \mathcal{F}_n\right]$ is uniformly integrable (with respect to μ).*

The proof of this lemma is a special case of part of the proof of Lévy's theorem in [Shiryaev, 1996]. We repeat it here for completeness.

Proof. For any constants $b, c > 0$

$$\begin{aligned}
\int_{X_n > b} X_n \, \mathrm{d}\mu &= \int_{X_n > b} p \, \mathrm{d}\mu \\
&\leq \int_{X_n > b, p \leq c} p \, \mathrm{d}\mu + \int_{X_n > b, p > c} p \, \mathrm{d}\mu \\
&\leq c \cdot \mu\left(X_n > b\right) + \int_{p>c} p \, \mathrm{d}\mu \\
&\overset{(*)}{\leq} \frac{c}{b} \, \mathbf{E}[X_n] + \int_{p>c} p \, \mathrm{d}\mu = \frac{c}{b} + \int_{p>c} p \, \mathrm{d}\mu,
\end{aligned}$$

in which the inequality marked by $(*)$ is Markov's. Consequently

$$\begin{aligned}
\lim_{b\to\infty} \sup_n \int_{X_n > b} |X_n| \, \mathrm{d}\mu &= \lim_{c\to\infty} \lim_{b\to\infty} \sup_n \int_{X_n > b} |X_n| \, \mathrm{d}\mu \\
&\leq \lim_{c\to\infty} \lim_{b\to\infty} \frac{c}{b} + \lim_{c\to\infty} \int_{p>c} p \, \mathrm{d}\mu = 0,
\end{aligned}$$

which proves the lemma. □

Proof of Theorem 6.18. The data processing inequality implies that $D_\alpha(P_{|\mathcal{F}_n} \| Q_{|\mathcal{F}_n}) \leq D_\alpha(P \| Q)$ for all n. We therefore only need to show that $\lim_{n\to\infty} D_\alpha(P_{|\mathcal{F}_n} \| Q_{|\mathcal{F}_n}) \geq D_\alpha(P_{|\mathcal{F}_\infty} \| Q_{|\mathcal{F}_\infty})$.

To this end, assume without loss of generality that $\mathcal{F} = \mathcal{F}_\infty$ and that μ is a probability distribution (i.e. $\mu = (P+Q)/2$). Let $X_n = \mathbf{E}\left[p \middle| \mathcal{F}_n\right]$ and $Y_n = \mathbf{E}\left[q \middle| \mathcal{F}_n\right]$, and define the distributions $\tilde{P}_n$ and $\tilde{Q}_n$ on $(\mathcal{X}, \mathcal{F})$ by

$$\tilde{P}_n(A) = \int_A X_n \,\mathrm{d}\mu, \quad \tilde{Q}_n(A) = \int_A Y_n \,\mathrm{d}\mu \qquad (A \in \mathcal{F}),$$

such that, by the Radon-Nikodým theorem and Proposition 6.1, $\frac{\mathrm{d}\tilde{P}_n}{\mathrm{d}\mu} = X_n = \frac{\mathrm{d}P_{|\mathcal{F}_n}}{\mathrm{d}\mu_{|\mathcal{F}_n}}$ and $\frac{\mathrm{d}\tilde{Q}_n}{\mathrm{d}\mu} = Y_n = \frac{\mathrm{d}Q_{|\mathcal{F}_n}}{\mathrm{d}\mu_{|\mathcal{F}_n}}$ (μ-a.s.) It follows that

$$D_\alpha(\tilde{P}_n \| \tilde{Q}_n) = D_\alpha(P_{|\mathcal{F}_n} \| Q_{|\mathcal{F}_n})$$

for $0 < \alpha < \infty$ and therefore by continuity also for $\alpha = \infty$. We will proceed to show that $(\tilde{P}_n, \tilde{Q}_n) \to (P, Q)$ in the τ-topology. By lower semi-continuity of Rényi divergence this implies that $\lim_{n\to\infty} D_\alpha(\tilde{P}_n \| \tilde{Q}_n) \geq D_\alpha(P\|Q)$, from which the theorem follows.

By Lévy's theorem [Shiryaev, 1996], $\lim_{n\to\infty} X_n = p$ (μ-a.s.) Hence uniform integrability of the family $\{X_n\}$ (by Lemma 6.4) implies that for any $A \in \mathcal{F}$

$$\lim_{n\to\infty} \tilde{P}_n(A) = \lim_{n\to\infty} \int_A X_n \,\mathrm{d}\mu = \int_A p \,\mathrm{d}\mu = P(A)$$

[Shiryaev, 1996, Thm. 5, p. 189]. Similarly $\lim_{n\to\infty} \tilde{Q}_n(A) = Q(A)$, so we find that $(\tilde{P}_n, \tilde{Q}_n) \to (P, Q)$, which completes the proof. □

Theorem 6.19 (Decreasing)**.** *Let $\mathcal{F} \supseteq \mathcal{F}_1 \supseteq \mathcal{F}_2 \supseteq \cdots$ be a nonincreasing family of σ-algebras, and let $\mathcal{F}_\infty = \bigcap_{n=1}^\infty \mathcal{F}_n$ be the largest σ-algebra contained in all of them. Let $0 \leq \alpha < \infty$. If $0 \leq \alpha < 1$ or there exists an m such that $D_\alpha(P_{|\mathcal{F}_m} \| Q_{|\mathcal{F}_m}) < \infty$, then*

$$\lim_{n\to\infty} D_\alpha(P_{|\mathcal{F}_n} \| Q_{|\mathcal{F}_n}) = D_\alpha(P_{|\mathcal{F}_\infty} \| Q_{|\mathcal{F}_\infty}).$$

Lemma 6.5. *Let $\mathcal{F} \supseteq \mathcal{F}_1 \supseteq \mathcal{F}_2 \supseteq \cdots$ be a nonincreasing family of σ-algebras. Let $0 < \alpha < \infty$, $p_n = \frac{\mathrm{d}P_{|\mathcal{F}_n}}{\mathrm{d}\mu_{|\mathcal{F}_n}}, q_n = \frac{\mathrm{d}Q_{|\mathcal{F}_n}}{\mathrm{d}\mu_{|\mathcal{F}_n}}$ and $X_n = f\left(\frac{p_n}{q_n}\right)$, where $f(x) = x^\alpha$ if $\alpha \neq 1$ and $f(x) = x\log x + e^{-1}$ if $\alpha = 1$. If $0 < \alpha < 1$, or $\mathbf{E}_Q[X_1] < \infty$ and $P \ll Q$, then the family $\{X_n\}_{n\geq 1}$ is uniformly integrable (with respect to Q).*

Proof. Suppose first that $0 < \alpha < 1$. Then for any $b > 0$

$$\begin{aligned}\int_{X_n>b} X_n \,\mathrm{d}Q &\le \int_{X_n>b} X_n \left(\frac{X_n}{b}\right)^{(1-\alpha)/\alpha} \mathrm{d}Q \\ &\le b^{-(1-\alpha)/\alpha} \int X_n^{1/\alpha} \,\mathrm{d}Q \le b^{-(1-\alpha)/\alpha},\end{aligned}$$

and, as $X_n \ge 0$, $\lim_{b\to\infty} \sup_n \int_{|X_n|>b} |X_n| \,\mathrm{d}Q = 0$, which was to be shown.

Alternatively, suppose that $1 \le \alpha < \infty$ and assume without loss of generality that $\mathcal{F} = \mathcal{F}_1$. Then $\frac{p_n}{q_n} = \frac{\mathrm{d}P|_{\mathcal{F}_n}}{\mathrm{d}Q|_{\mathcal{F}_n}}$ (Q-a.s.) and hence by Proposition 6.1 and Jensen's inequality for conditional expectations

$$X_n = f\left(\mathbf{E}\left[\frac{\mathrm{d}P}{\mathrm{d}Q}\middle|\,\mathcal{F}_n\right]\right) \le \mathbf{E}\left[f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\middle|\,\mathcal{F}_n\right] = \mathbf{E}\left[X_1|\,\mathcal{F}_n\right] \quad (Q\text{-a.s.})$$

As $\min_x x \log x = -e^{-1}$, it follows that $X_n \ge 0$ and for any $b, c > 0$

$$\begin{aligned}\int_{|X_n|>b} |X_n| \,\mathrm{d}Q &= \int_{X_n>b} X_n \,\mathrm{d}Q \\ &\le \int_{X_n>b} \mathbf{E}\left[X_1|\,\mathcal{F}_n\right] \mathrm{d}Q = \int_{X_n>b} X_1 \,\mathrm{d}Q \\ &= \int_{X_n>b, X_1\le c} X_1 \,\mathrm{d}Q + \int_{X_n>b, X_1>c} X_1 \,\mathrm{d}Q \\ &\le c \cdot Q(X_n > b) + \int_{X_1>c} X_1 \,\mathrm{d}Q \\ &\le \frac{c}{b}\,\mathbf{E}_Q[X_n] + \int_{X_1>c} X_1 \,\mathrm{d}Q \\ &\le \frac{c}{b}\,\mathbf{E}_Q[X_1] + \int_{X_1>c} X_1 \,\mathrm{d}Q,\end{aligned}$$

where $\mathbf{E}_Q[X_n] \le \mathbf{E}_Q[X_1]$ in the last inequality follows from the data processing inequality. Consequently,

$$\begin{aligned}\lim_{b\to\infty} \sup_n \int_{|X_n|>b} |X_n| \,\mathrm{d}Q &= \lim_{c\to\infty} \lim_{b\to\infty} \sup_n \int_{|X_n|>b} |X_n| \,\mathrm{d}Q \\ &\le \lim_{c\to\infty} \lim_{b\to\infty} \frac{c}{b}\,\mathbf{E}_Q[X_1] + \lim_{c\to\infty} \int_{X_1>c} X_1 \,\mathrm{d}Q = 0,\end{aligned}$$

and the lemma follows. □

Proof of Theorem 6.19. First suppose that $\alpha > 0$. For $n = 1, 2, \ldots, \infty$, let $p_n = \frac{\mathrm{d}P_{|\mathcal{F}_n}}{\mathrm{d}\mu_{|\mathcal{F}_n}}, q_n = \frac{\mathrm{d}Q_{|\mathcal{F}_n}}{\mathrm{d}\mu_{|\mathcal{F}_n}}$ and $X_n = f\left(\frac{p_n}{q_n}\right)$ with $f(x) = x^\alpha$ if $\alpha \neq 1$ and $f(x) = x \log x + e^{-1}$ if $\alpha = 1$, as in Lemma 6.5. If $\alpha \geq 1$, then assume without loss of generality that $\mathcal{F} = \mathcal{F}_1$ and $m = 1$, such that $D_\alpha(P_{|\mathcal{F}_m} \| Q_{|\mathcal{F}_m}) < \infty$ implies $P \ll Q$. Now, for any $\alpha > 0$, it is sufficient to show that

$$\mathbf{E}_Q[X_n] \to \mathbf{E}_Q[X_\infty]. \tag{6.20}$$

By Proposition 6.1, $p_n = \mathbf{E}_\mu\left[p \middle| \mathcal{F}_n\right]$ and $q_n = \mathbf{E}_\mu\left[q \middle| \mathcal{F}_n\right]$. Therefore by a version of Lévy's theorem for decreasing sequences of σ-algebras [Kallenberg, 1997, Theorem 6.23],

$$\begin{aligned} p_n = \mathbf{E}_\mu\left[p \middle| \mathcal{F}_n\right] &\to \mathbf{E}_\mu\left[p \middle| \mathcal{F}_\infty\right] = p_\infty, \\ q_n = \mathbf{E}_\mu\left[q \middle| \mathcal{F}_n\right] &\to \mathbf{E}_\mu\left[q \middle| \mathcal{F}_\infty\right] = q_\infty, \end{aligned} \qquad (\mu\text{-a.s.})$$

and hence $X_n \to X_\infty$ (μ-a.s. and therefore Q-a.s.)

If $0 < \alpha < 1$, then

$$\mathbf{E}_Q[X_n] = \mathbf{E}_\mu\left[p_n^\alpha q_n^{1-\alpha}\right] \leq \mathbf{E}_\mu\left[\alpha p_n + (1-\alpha) q_n\right] = 1 < \infty.$$

And if $\alpha \geq 1$, then from the data processing inequality we get that $D_\alpha(P_{|\mathcal{F}_n} \| Q_{|\mathcal{F}_n}) < \infty$ for all n, which implies that also in this case $\mathbf{E}_Q[X_n] < \infty$. Hence uniform integrability (by Lemma 6.5) of the family of nonnegative random variables $\{X_n\}$ implies (6.20) [Shiryaev, 1996, Thm. 5, p. 189], and the theorem follows for $\alpha > 0$. The remaining case, $\alpha = 0$, is proved by

$$\begin{aligned} \lim_{n\to\infty} D_0(P_{|\mathcal{F}_n} \| Q_{|\mathcal{F}_n}) &= \inf_n \inf_{\alpha>0} D_\alpha(P_{|\mathcal{F}_n} \| Q_{|\mathcal{F}_n}) \\ &= \inf_{\alpha>0} \inf_n D_\alpha(P_{|\mathcal{F}_n} \| Q_{|\mathcal{F}_n}) = \inf_{\alpha>0} D_\alpha(P_{|\mathcal{F}_\infty} \| Q_{|\mathcal{F}_\infty}) \\ &= D_0(P_{|\mathcal{F}_\infty} \| Q_{|\mathcal{F}_\infty}). \end{aligned}$$

□

6.5.6 Distributions on Sequences

Suppose $(\mathcal{X}^\infty, \mathcal{F}^\infty)$ is the *direct product* of an infinite sequence of measurable spaces $(\mathcal{X}_1, \mathcal{F}_1), (\mathcal{X}_2, \mathcal{F}_2), \ldots$ That is, $\mathcal{X}^\infty = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots$ and $\mathcal{F}^\infty$ is the smallest σ-algebra containing all the *cylinder sets*

$$S_n(A) = \{x^\infty \in \mathcal{X}^\infty \mid x_1, \ldots, x_n \in A\}, \qquad A \in \mathcal{F}^n,$$

for $n = 1, 2, \ldots$, where $\mathcal{F}^n = \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_n$. Then a sequence of probability distributions $P^1, P^2, \ldots$, where P^n is a distribution on $\mathcal{X}^n = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, is called *consistent* if

$$P^{n+1}(A \times \mathcal{X}_{n+1}) = P^n(A), \qquad A \in \mathcal{F}^n.$$

For any such consistent sequence there exists a distribution P^∞ on $(\mathcal{X}^\infty, \mathcal{F}^\infty)$ such that its marginal distribution on $\mathcal{X}^n$ is P^n, in the sense that

$$P^\infty(S_n(A)) = P^n(A), \qquad A \in \mathcal{F}^n.$$

If $P^1, P^2, \ldots$ and $Q^1, Q^2, \ldots$ are two consistent sequences of probability distributions, then it is natural to ask whether the Rényi divergence $D_\alpha(P^n \| Q^n)$ converges to $D_\alpha(P^\infty \| Q^\infty)$. The following theorem shows that it does for $\alpha > 0$.

Theorem 6.20. *Let $P^1, P^2, \ldots$ and $Q^1, Q^2, \ldots$ be consistent sequences of probability distributions on $(\mathcal{X}^1, \mathcal{F}^1), (\mathcal{X}^2, \mathcal{F}^2), \ldots$, where, for $n = 1, \ldots, \infty$, $(\mathcal{X}^n, \mathcal{F}^n)$ is the direct product of the first n measurable spaces in the infinite sequence $(\mathcal{X}_1, \mathcal{F}_1), (\mathcal{X}_2, \mathcal{F}_2), \ldots$ Then for any $0 < \alpha \le \infty$*

$$D_\alpha(P^n | Q^n) \to D_\alpha(P^\infty | Q^\infty)$$

as $n \to \infty$.

Proof. Let $\mathcal{G}^n = \{S_n(A) \mid A \in \mathcal{F}^n\}$. Then

$$D_\alpha(P^n | Q^n) = D_\alpha(P^\infty_{|\mathcal{G}^n} \| Q^\infty_{|\mathcal{G}^n}) \to D_\alpha(P^\infty \| Q^\infty)$$

by Theorem 6.18. □

As a special case, we find that finite additivity of Rényi divergence, which is easy to verify, extends to countable additivity:

Theorem 6.21 (Additivity)**.** *For $n = 1, 2, \ldots$, let (P_n, Q_n) be pairs of probability distributions on measurable spaces $(\mathcal{X}_n, \mathcal{F}_n)$. Then for any $0 \le \alpha \le \infty$ and any $N \in \{1, 2, \ldots\}$*

$$\sum_{n=1}^{N} D_\alpha(P_n \| Q_n) = D_\alpha(P_1 \times \cdots \times P_N \| Q_1 \times \cdots \times Q_N), \tag{6.21}$$

and, except for $\alpha = 0$, also

$$\sum_{n=1}^{\infty} D_\alpha(P_n \| Q_n) = D_\alpha(P_1 \times P_2 \times \cdots \| Q_1 \times Q_2 \times \cdots). \tag{6.22}$$

Countable additivity as in (6.22) does not hold for $\alpha = 0$. A counterexample is given following Example 6.1 below.

Proof. For simple orders α, (6.21) follows from independence of P_n and Q_n between different n, which implies that

$$\prod_{n=1}^{N} \int \left(\frac{\mathrm{d}Q_n}{\mathrm{d}P_n}\right)^{1-\alpha} \mathrm{d}P_n = \int \left(\frac{\mathrm{d}\prod_{n=1}^{N} Q_n}{\mathrm{d}\prod_{n=1}^{N} P_n}\right)^{1-\alpha} \mathrm{d}\prod_{n=1}^{N} P_n.$$

As N is finite, this extends to the extended orders by continuity in α. Finally, (6.22) follows from Theorem 6.20 by observing that the sequences $P^N = P_1 \times \cdots \times P_N$ and $Q^N = Q_1 \times \cdots \times Q_N$, for $N = 1, 2, \ldots$, are consistent. □

6.5.7 Absolute Continuity and Mutual Singularity

Shiryaev [1996, pp. 366,370] relates Hellinger integrals to absolute continuity and mutual singularity of probability distributions. His results may also be expressed in terms of Rényi divergence. They then follow from the observations that $D_0(P\|Q) = 0$ if and only if Q is absolutely continuous with respect to P and that $D_0(P\|Q) = \infty$ if and only if P and Q are mutually singular, together with right-continuity of $D_\alpha(P\|Q)$ in α at $\alpha = 0$.

Theorem 6.22 ([Shiryaev, 1996, Theorem 2, p. 366])**.** *The following conditions are equivalent:*

(a) $Q \ll P$,

(b) $Q(p > 0) = 1$,

(c) $D_0(P\|Q) = 0$,

(d) $\lim_{\alpha \downarrow 0} D_\alpha(P\|Q) = 0$.

Proof. Clearly (b) is equivalent to $Q(p = 0) = 0$, which is equivalent to (a). The other cases follow by

$$\lim_{\alpha \downarrow 0} D_\alpha(P\|Q) = D_0(P\|Q) = -\log Q(p > 0).$$ □

Theorem 6.23 ([Shiryaev, 1996, Theorem 3, p. 366])**.** *The following conditions are equivalent:*

(a) $Q \perp P$,

(b) $Q(p > 0) = 0$,

(c) $D_\alpha(P\|Q) = \infty$ *for some* $0 \le \alpha < 1$,

(d) $D_\alpha(P\|Q) = \infty$ *for all* $\alpha \ge 0$.

Proof. Equivalence of (a),(b) and $D_0(P\|Q) = \infty$ follows from definitions. Equivalence of $D_0(P\|Q) = \infty$ and (d) follows from the fact the Rényi divergence is continuous on $[0, 1]$ and nondecreasing in α. Finally, (c) for some $0 < \alpha < 1$ is equivalent to

$$\int p^\alpha q^{1-\alpha} \, \mathrm{d}\mu = 0,$$

which holds if and only if $pq = 0$ (μ-a.s.). It follows that in this case (c) is equivalent to (a). □

These properties give a convenient mathematical tool to prove absolute continuity or mutual singularity of infinite product distributions, as illustrated by the following proof by Shiryaev [1996] of the *Gaussian dichotomy* [Feldman, 1958, Hájek, 1958, Thelen, 1989].

Example 6.1 (Gaussian Dichotomy)**.** Let $P = P_1 \times P_2 \times \cdots$ and $Q = Q_1 \times Q_2 \times \cdots$, where P_n and Q_n are Gaussian distributions with densities

$$p_n(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu_n)^2}, \quad q_n(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\nu_n)^2}.$$

Then for simple orders α

$$D_\alpha(P_n\|Q_n) = \frac{1}{2}\alpha(\mu_n - \nu_n)^2,$$

and by additivity

$$D_\alpha(P\|Q) = \frac{1}{2}\alpha \sum_{n=1}^{\infty} (\mu_n - \nu_n)^2.$$

Consequently, by Theorems 6.22 and 6.23:

$$Q \ll P \quad \Leftrightarrow \quad P \ll Q \quad \Leftrightarrow \quad \sum_{n=1}^{\infty} (\mu_n - \nu_n)^2 < \infty,$$

$$Q \perp P \quad \Leftrightarrow \quad \sum_{n=1}^{\infty} (\mu_n - \nu_n)^2 = \infty.$$

The observation that P and Q are either equivalent (both $P \ll Q$ and $Q \ll P$) or mutually singular is called the *Gaussian dichotomy*.

Example 6.1 shows that countable additivity does not hold for $\alpha = 0$: if $\sum_{n=1}^{\infty}(\mu_n - \nu_n)^2 = \infty$, then $\sum_{n=1}^{N} D_0(P_n\|Q_n) = 0$ for all N, while $D_0(P\|Q) = \infty$. In light of the proof of Theorem 6.21 this also provides a counterexample to (6.19) for $\alpha = 0$.

The Gaussian dichotomy raises the question of whether the same dichotomy holds for other product distributions. Let $P \sim Q$ denote that P and Q are *equivalent* (both $P \ll Q$ and $Q \ll P$). Suppose that $P = P_1 \times P_2 \times \cdots$ and $Q = Q_1 \times Q_2 \times \cdots$, where P_n and Q_n are arbitrary distributions on arbitrary measurable spaces. Then if $P_n \not\sim Q_n$ for some n, P and Q are not equivalent either. The question is therefore answered by the following theorem:

Theorem 6.24 (Kakutani's Dichotomy)**.** *Let $0 < \alpha < 1$ and let $P = P_1 \times P_2 \times \cdots$ and $Q = Q_1 \times Q_2 \times \cdots$, where P_n and Q_n are distributions on arbitrary measurable spaces such that $P_n \sim Q_n$. Then*

$$Q \sim P \quad \Leftrightarrow \quad \sum_{n=1}^{\infty} D_\alpha(P_n\|Q_n) < \infty,$$

$$Q \perp P \quad \Leftrightarrow \quad \sum_{n=1}^{\infty} D_\alpha(P_n\|Q_n) = \infty.$$

Proof. If $\sum_{n=1}^{\infty} D_\alpha(P_n\|Q_n) = \infty$, then $D_\alpha(P\|Q) = \infty$ and $Q \perp P$ follows by Theorem 6.23. On the other hand, if $\sum_{n=1}^{\infty} D_\alpha(P_n\|Q_n) < \infty$, then for every $\varepsilon > 0$ there exists an N such that

$$\sum_{n=N+1}^{\infty} D_\alpha(P_n\|Q_n) \le \varepsilon,$$

and consequently by additivity and monotonicity in α:

$$\begin{aligned} D_0(P\|Q) &= \lim_{\alpha\downarrow 0} D_\alpha(P\|Q) \\ &\le \lim_{\alpha\downarrow 0} D_\alpha(P_1 \times \cdots \times P_N\|Q_1 \times \cdots \times Q_N) + \varepsilon = \varepsilon. \end{aligned}$$

As this holds for any $\varepsilon > 0$, $D_0(P\|Q)$ must equal 0, and, by Theorem 6.22, $Q \ll P$. As $Q \ll P$ implies $Q \not\perp P$, Theorem 6.23 implies that $D_\alpha(Q\|P) < \infty$, and by repeating the argument with the roles of P and Q reversed we find that also $P \ll Q$, which completes the proof. □

Theorem 6.24 (with $\alpha = 1/2$) is equivalent to a classical result by Kakutani [1948], which was stated in terms of Hellinger integrals rather than Rényi divergence, and according to Gibbs and Su [2002] might be responsible for popularising Hellinger integrals. Kakutani's result is related to the amount of information that a sequence of observations contains about the parameter of a statistical model [Rényi, 1967]. Our simple proof in terms of Rényi divergence illustrates that whether $P \sim Q$ or $P \perp Q$ really depends on whether $D_\alpha(P_{N+1} \times P_{N+2} \times \cdots \| Q_{N+1} \times Q_{N+2} \times \cdots) \to 0$ as $N \to \infty$.

6.6 Applications and Further References

Rényi divergence comes up in many settings, most of which are related to hypothesis testing. We give a unified overview and references for further reading.

6.6.1 Hypothesis Testing

Rényi divergence appears in bounds on the error probabilities when testing a probabilistic hypothesis Q against an alternative P [Nemetz, 1974, Rached et al., 2001] and in classification problems [Ben-Bassat and Raviv, 1978]. Csiszár [1995] provides the following explanation: let $Z_i = \log P(X_i)/Q(X_i)$, where $X_1, X_2, \ldots$ are discrete random variables that are distributed independently and identically (i.i.d.) according to a Q. Since large deviation theory involves the *moment generating function* $M(\alpha) = \mathbf{E}_Q[e^{\alpha Z_i}]$ (see [Chernoff, 1952]), the observation that

$$(1-\alpha)D_\alpha(P\|Q) = -\log M(\alpha) \qquad \text{for simple orders } \alpha \tag{6.23}$$

as long as $0 < \alpha < 1$ or $P \ll Q$, explains the appearance of Rényi divergence in hypothesis testing. Using the connection (6.23) and well-known properties of $\log M(\alpha)$, Grünwald [2007, Section 19.6] finds that, under regularity conditions, $(1-\alpha)D_\alpha(P\|Q)$ is strictly concave in α if $P \neq Q$. The same connection is exploited in the proof of Lemma 5.2 from the previous chapter. The analysis there takes the hypothesis code lengths (or priors) into account, which is necessary to deal with an infinite number of hypotheses.

To obtain asymptotically tight bounds, Chernoff uses the supremum of $-\log M(\alpha)$ over $\alpha \in (0,1)$, which is called the *Chernoff information*.

The following theorem relates this quantity to an information divergence involving the distribution P_α with density

$$p_\alpha = \frac{p^\alpha q^{1-\alpha}}{\int p^\alpha q^{1-\alpha} \, \mathrm{d}\mu}, \tag{6.24}$$

which is well defined if and only if $0 < \int p^\alpha q^{1-\alpha} \, \mathrm{d}\mu < \infty$.

Theorem 6.25. *Let P and Q be distributions, and let P_α denote the distribution with density (6.24). If there exists a simple order α^* such that P_{α^*} is well defined, $D(P_{\alpha^*}\|P) = D(P_{\alpha^*}\|Q)$ and either $0 < \alpha^* < 1$ or $D(P_{\alpha^*}\|P) < \infty$, then*

$$\sup_\alpha \, (1-\alpha) D_\alpha(P\|Q) = D(P_{\alpha^*}\|P),$$

where the supremum is over simple orders α.

The same connection between Chernoff information and $D(P_{\alpha^*}\|P)$ is discussed by Cover and Thomas [1991, Section 12.9], but our proof is different. As an intermediate step we use the following lemma, which is interesting in itself, because it gives an interpretation of Rényi divergence as a trade-off between two information divergences:

Lemma 6.6. *Let P and Q be probability distributions and let α be a simple order. Then*

$$(1-\alpha) D_\alpha(P\|Q) = \inf_R \Big\{ \alpha D(R\|P) + (1-\alpha) D(R\|Q) \Big\}, \tag{6.25}$$

with the convention that $\alpha D(R\|P) + (1-\alpha) D(R\|Q) = \infty$ if it would otherwise be undefined. Moreover, if the distribution P_α with density (6.24) is well defined and $0 < \alpha < 1$ or $D(P_\alpha\|P) < \infty$, then the infimum is uniquely achieved by $R = P_\alpha$.

Proof. First suppose that P_α is well defined or, equivalently, that $D_\alpha(P\|Q) < \infty$. Then for $0 < \alpha < 1$ or $D(R\|P) < \infty$, we have

$$\alpha D(R\|P) + (1-\alpha) D(R\|Q) = D(R\|P_\alpha) - \log \int p^\alpha q^{1-\alpha} \, \mathrm{d}\mu.$$

Hence, if $0 < \alpha < 1$ or $D(P_\alpha\|P) < \infty$, the infimum over R is uniquely achieved by $R = P_\alpha$, for which it equals $(1-\alpha) D_\alpha(P\|Q)$ as required. If, on the other hand, $\alpha > 1$ and $D(P_\alpha\|P) = \infty$, then we still have

$$\inf_R \Big\{ \alpha D(R\|P) + (1-\alpha) D(R\|Q) \Big\} \geq (1-\alpha) D_\alpha(P\|Q). \tag{6.26}$$

Secondly, suppose $0 < \alpha < 1$ and $D_\alpha(P\|Q) = \infty$. Then $P \perp Q$, and consequently either $D(R\|P) = \infty$ or $D(R\|Q) = \infty$ for all R, so that (6.25) holds.

Next, consider the case that $\alpha > 1$ and $P \not\ll Q$. Then $D_\alpha(P\|Q) = \infty$ and the infimum over R is achieved by $R = P$, for which it equals $-\infty$, so that (6.25) holds.

Finally, we prove (6.25) for the remaining cases: $\alpha > 1, P \ll Q$ and either: (1) $D_\alpha(P\|Q) < \infty$, but $D(P_\alpha\|P) = \infty$; or (2) $D_\alpha(P\|Q) = \infty$. To this end, let $P_c = P(\cdot \mid p \le cq)$ for all c that are sufficiently large that $P(p \le cq) > 0$. The reader may verify that $D_\alpha(P_c\|Q) < \infty$ and $D(S\|P_c) < \infty$ for $s = p_c^\alpha q^{1-\alpha} / \int p_c^\alpha q^{1-\alpha}\, d\mu$, so that we have already proved that (6.25) holds if P is replaced by P_c. Hence, observing that for all R

$$D(R\|P_c) = \begin{cases} \infty & \text{if } R \not\ll P_c, \\ D(R\|P) + \log P(p \le pc) & \text{otherwise,} \end{cases}$$

we find that

$$\begin{aligned} &\inf_R \Big\{ \alpha D(R\|P) + (1-\alpha) D(R\|Q) \Big\} \\ &\quad \le \limsup_{c\to\infty} \left(-\alpha \log P(p \le cq) + \inf_R \Big\{ \alpha D(R\|P_c) + (1-\alpha) D(R\|Q) \Big\} \right) \\ &\quad \le \limsup_{c\to\infty} (1-\alpha) D_\alpha(P_c\|Q) \le (1-\alpha) D_\alpha(P\|Q), \end{aligned}$$

where the last inequality follows by lower semi-continuity of D_α (Theorem 6.15). In Case 2, (6.25) follows immediately. In Case 1, (6.25) follows by combining this inequality with its converse (6.26). □

Theorem 6.25 follows almost immediately from Lemma 6.6:

Proof of Theorem 6.25. Let

$$f(\alpha, R) = \alpha D(R\|P) + (1-\alpha) D(R\|Q).$$

By Lemma 6.6

$$\sup_\alpha (1-\alpha) D_\alpha(P\|Q) = \sup_\alpha \inf_R f(\alpha, R)$$

and f has a saddle-point at $\alpha = \alpha^*$ and $R = P_{\alpha^*}$. This implies that

$$\sup_\alpha \inf_R f(\alpha, R) = f(\alpha^*, P_{\alpha^*})$$

(see for example [Rockafellar, 1970, Lemma 36.2]), from which the theorem follows. □

6.6.2 Further References

We discuss some remaining properties of Rényi divergence that are not directly related to hypothesis testing.

Multiple Source Adaptation *Source adaptation* is a statistical learning problem in which test data are assumed to come from a related, but slightly different source than training data, such that their distribution is slightly different. Mansour et al. [2009] study source adaptation problems with multiple training sources, each having their own distribution. The starting point of their analysis is a lemma showing that the expectation of a bounded random variable cannot increase much if its distribution is replaced by another distribution that is close in Rényi divergence. They essentially prove the following:

Lemma 6.7. *Let P and Q be distributions on the same sample space, and let $X \leq b$ be a random variable. Then for any $\alpha > 1$*

$$\log \mathbf{E}_{X\sim P}[X] \leq \frac{\alpha-1}{\alpha} \log \mathbf{E}_{X\sim Q}[X] + \frac{\alpha-1}{\alpha} D_\alpha(P\|Q) + \frac{1}{\alpha} \log b.$$

They also prove a property that is similar to the triangle inequality, except for an increase in α:

Lemma 6.8. *For any distributions P, Q and R and any $\alpha > 1$*

$$D_\alpha(P\|R) \leq D_{2\alpha}(P\|Q) + D_{2\alpha-1}(Q\|R).$$

Guessing Rényi's entropy and divergence are also related to the moments of *guessing functions*, which determine an order for guessing the values of a random variable in sequential decoding, and thereby the computational complexity of the corresponding decoder [Arikan, 1996, Van Erven and Harremoes, 2010].

Taylor Approximation for Parametric Models Suppose $\mathcal{M} = \{P_\theta \mid \theta \in \Theta \subseteq \mathbb{R}\}$ is a parametric statistical model. Then it is well known

that, for sufficiently regular parametrisations, a second order Taylor approximation of $D(P_\theta \| P_{\theta'})$ in θ' at θ in the interior of Θ yields

$$\lim_{\theta' \to \theta} \frac{1}{(\theta - \theta')^2} D(P_\theta \| P_{\theta'}) = \frac{1}{2} J(\theta),$$

where $J(\theta) = \mathbf{E}(\frac{\mathrm{d}}{\mathrm{d}\theta} \log p_\theta)^2$ denotes the *Fisher information* at θ (see e.g. [Cover and Thomas, 1991, Problem 12.7]). Haussler and Opper [1997] argue that this property generalises to

$$\lim_{\theta' \to \theta} \frac{1}{(\theta - \theta')^2} D_\alpha(P_\theta \| P_{\theta'}) = \frac{\alpha}{2} J(\theta)$$

for any $0 < \alpha < \infty$.

Ranking Images Hero et al. [2003] use estimated Rényi divergence on densities of feature vectors to rank images in a database by their Rényi divergence from a fixed reference image.

6.7 Conclusion

We have extended the definition of Rényi divergence from discrete to continuous spaces, and confirmed (by Theorem 6.2) that this is the appropriate generalisation. The most important properties of Rényi divergence were reviewed, and connections to absolute continuity of distributions and hypothesis testing were elaborated on.

Bibliography

J. Aczél and Z. Daróczy. *On Measures of Information and Their Characterizations*. Academic Press, 1975.

H. Akaike. A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika*, 66(2):237–242, 1979.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

D. Aldous and P. Diaconis. Strong uniform times and finite random walks. *Advances in Applied Mathematics*, 8:69–97, 1987.

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, series B*, 28(1):131–142, 1966.

F. J. Anscombe. The transformation of Poisson, binomial and negative-binomial data. *Biometrika*, 35:246–254, 1948.

E. Arikan. An inequality on guessing and its application to sequential decoding. *IEEE Transactions on Information Theory*, 42(1):99–105, 1996.

V. Balasubramanian. Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Computation*, 9:349–368, 1997.

Y. Baraud. Model selection for regression on a random design. *ESAIM: Probability and Statistics*, 6:127–146, 2002.

A. Barron and C. Sheu. Approximation of density functions by sequences of exponential families. *The Annals of Statistics*, 19(3):1347–1369, 1991.

A. Barron, Y. Yang, and B. Yu. Asymptotically optimal function estimation by minimum complexity criteria. In *Proceedings of the 1994 International Symposium on Information Theory*, page 38, Trondheim, Norway, 1994.

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

A. Barron, M. J. Schervish, and L. Wasserman. The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561, 1999.

A. R. Barron. *Logically Smooth Density Estimation*. PhD thesis, Stanford University, 1985.

A. R. Barron. Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 6*, pages 27–52. Oxford University Press, 1998.

A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4):1034–1054, 1991.

M. Ben-Bassat and J. Raviv. Renyi's entropy and the probability of error. *IEEE Transactions on Information Theory*, 24(3):324–330, 1978.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley, 1994.

L. Birgé. Model selection for Gaussian regression with random design. *Bernoulli*, 10(6):1039–1051, 2004.

L. Birgé. On estimating a density using Hellinger distance and some other strange facts. *Probability Theory and Related Fields*, 71:271–291, 1986.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

O. Bousquet and M. K. Warmuth. Tracking a small set of experts by mixing past posteriors. *Journal of Machine Learning Research*, 3:363–396, 2002.

K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference*. Springer-Verlag, second edition, 2002.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

B. Clarke. Online forecasting proposal. Technical report, University of Dortmund, 1997. Sonderforschungsbereich 475.

B. Clarke and A. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41: 37–60, 1994.

B. S. Clarke and A. R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, May 1990.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

I. Csiszár. Generalized cutoff rates and Rényi's information measures. *IEEE Transactions on Information Theory*, 41(1):26–34, 1995.

I. Csiszár and P. C. Shields. The consistency of the BIC Markov order estimator. *The Annals of Statistics*, 28(6):1601–1619, 2000.

A. Dawid. Prequential data analysis. In M. Gosh and P. Pathak, editors, *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, volume 17 of *IMS Lecture Notes*, pages 113–125, 1992a.

A. P. Dawid. Prequential analysis, stochastic complexity and Bayesian inference. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 109–125. Oxford University Press, 1992b.

A. P. Dawid. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A*, 147, Part 2:278–292, 1984.

X. de Luna and K. Skouras. Choosing a model selection strategy. *Scandinavian Journal of Statistics*, 30:113–128, 2003.

S. de Rooij and P. Grünwald. *Handbook of the Philosophy of Statistics*, volume 7, chapter VII-B. Elsevier, 2010.

R. DeVore and G. Lorentz. *Constructive Approximation*. Springer, 1993.

P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *The Annals of Statistics*, 14(1):1–26, 1986.

P. Domingos. The role of Occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery*, 3(4):409–425, 1999.

D. Donoho and I. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

J. Feldman. Equivalence and perpendicularity of Gaussian processes. *Pacific Journal of Mathematics*, 8(4):699–708, 1958.

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222:309–368, 1922.

M. Forster. The new science of simplicity. In A. Zellner, H. Keuzenkamp, and M. McAleer, editors, *Simplicity, Inference and Modelling*, pages 83–117. Cambridge University Press, Cambridge, 2001.

D. Foster and E. George. The risk inflation criterion for multiple regression. *Annals of Statistics*, 22:1947–1975, 1994.

M. Freeman and J. Tukey. Transformations related to the angular and the square root. *Annals of Mathematical Statistics*, 21:607–611, 1950.

S. Ghosal, J. K. Ghosh, and A. van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531, 2000.

S. Ghosal, J. Lember, and A. van der Vaart. Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics*, 2:63–89, 2008.

A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70:419–435, 2002.

N. Goodman. *Fact, Fiction, and Forecast*. Harvard University Press, 1955.

U. Grenander. *Abstract Inference*. Wiley, 1981.

P. Grünwald and W. Kotłowski. Prequential plug-in codes that achieve optimal redundancy rates even if the model is wrong. *IEEE International Symposium on Information Theory (ISIT)*, pages 1383–1387, 2010.

P. D. Grünwald. *The Minimum Description Length Principle*. The MIT Press, 2007.

P. D. Grünwald and S. de Rooij. Asymptotic log-loss of prequential maximum likelihood codes. In *Learning Theory: 18th Annual Conference on Learning Theory (COLT 2005)*, volume 3559 of *Lecture Notes in Computer Science*, pages 652–667, June 2005.

J. Hájek. On a property of normal distributions of any stochastic process. *Czechoslovak Mathematical Journal*, 8(4):610–618, 1958. In Russian with English summary.

J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.

M. Hansen and B. Yu. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96 (454):746–774, 2001.

M. Hansen and B. Yu. Minimum description length model selection criteria for generalized linear models. In *Science and Statistics: Festschrift for Terry Speed*, volume 40 of *IMS Lecture Notes – Monograph Series*. Institute for Mathematical Statistics, Hayward, CA, 2002.

P. Harremoës. Interpretations of Rényi entropies and divergences. *Physica A*, 365:57–62, 2006.

D. Haussler and M. Opper. Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, 25(6):2451–2492, 1997.

D. Haussler, J. Kivinen, and M. K. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44(5):1906–1925, 1998.

E. M. Hemerly and M. H. A. Davis. Strong consistency of the PLS criterion for order determination of autoregressive processes. *The Annals of Statistics*, 17(2):941–946, 1989.

M. Herbster and M. Warmuth. Tracking the best expert. *Machine Learning*, 32:151–178, 1998.

A. O. Hero, B. Ma, O. Michel, and J. D. Gorman. Alpha-divergence for classification, indexing and retrieval (revised). Technical Report CSPL-334, Communications and Signal Processing Laboratory, The University of Michigan, 2003.

S. Kakutani. On equivalence of infinite product measures. *The Annals of Mathematics*, 49(1):214–224, 1948.

O. Kallenberg. *Foundations of Modern Probability*. Springer, 1997.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, June 1995.

A. N. Kolmogorov. Complexity of algorithms and objective definition of randomness. In *Uspekhi Matematicheskikh Nauk*, volume 29, issue 4, page 155, 1974. English translation available in [Vereshchagin and Vitányi, 2004].

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

W. Koolen and S. de Rooij. Combining expert advice efficiently. In *Proceedings of the 21st Annual Conference on Computational Learning Theory (COLT)*, pages 275–286, 2008a.

W. Koolen and S. de Rooij. Combining expert advice efficiently. Available from arXiv.org, arXiv:0802.2015v2 [cs.LG], 2008b.

W. Kotłowski, P. Grünwald, and S. de Rooij. Following the flattened leader. In *Conference on Learning Theory (COLT)*, pages 106–118, 2010.

L. Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 1973.

K. Li. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics*, 15: 958–975, 1987.

M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, 2008.

F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10): 4394–4412, 2006.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Multiple source adaptation and the Rényi divergence. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI-09)*, pages 367–374, 2009.

C. Monteleoni and T. Jaakkola. Online learning of non-stationary sequences. In *Advances in Neural Information Processing Systems (NIPS)*, volume 16, Cambridge, MA, 2003. MIT Press.

T. Nemetz. On the α-divergence rate for Markov-dependent hypotheses. *Problems of Control and Information Theory*, 3(2):147–155, 1974.

K. R. Parthasarathy. *Probability Measures on Metric Spaces*. Probability and Mathematical Statistics. Academic Press, 1967.

J. Poland and M. Hutter. Asymptotics of discrete MDL for online prediction. *IEEE Transactions on Information Theory*, 51(11):3780–3795, 2005.

L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, issue 2, pages 257–285, 1989.

Z. Rached, F. Alajaji, and L. L. Campbell. Rényi's divergence and entropy rates for finite alphabet markov sources. *IEEE Transactions on Information Theory*, 47(4):1553–1561, 2001.

A. Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 547–561, 1961.

A. Rényi. On some basic problems of statistics from the point of view of information theory. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1: Statistics, pages 531–543, 1967.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

J. Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2):416–431, 1983.

J. Rissanen. Stochastic complexity and modeling. *The Annals of Statistics*, 14(3):1080–1100, 1986.

J. Rissanen. *Stochastic Complexity in Statistical Inquiry*, volume 15 of *Series in Computer Science*. World Scientific, 1989.

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

J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, IT-30(4):629–636, July 1984.

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

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

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

M. W. Seeger, S. M. Kakade, and D. P. Foster. Information consistency of nonparametric Gaussian process methods. *IEEE Transactions on Information Theory*, 54(5):2376–2382, 2008.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423,623–656, 1948.

R. Shibata. Asymptotic mean efficiency of a selection of regression variables. *Annals of the Institute of Statical Mathematics*, 35:415–423, 1983.

A. N. Shiryaev. *Probability*. Springer-Verlag, 1996.

Y. M. Shtar'kov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):175–186, 1987.

E. Sober. The contest between parsimony and likelihood. *Systematic Biology*, 4:644–653, 2004.

T. Speed and B. Yu. Model selection and prediction: Normal regression. *Annals of the Institute of Statical Mathematics*, 45(1):35–54, 1993.

M. Stone. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society B*, 39:44–47, 1977.

R. Sundaresan. A measure of discrimination and its geometric properties. In *IEEE International Symposium on Information Theory (ISIT)*, 2002.

R. Sundaresan. Guessing under source uncertainty with side information. In *IEEE International Symposium on Information Theory (ISIT)*, 2006.

E. Takimoto and M. Warmuth. The last-step minimax algorithm. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory (COLT 2000)*, pages 100–106, 2000.

B. J. Thelen. Fisher information and dichotomies in equivalence/contiguity. *The Annals of Probability*, 17(4):1664–1690, 1989.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

F. Topsøe. *Entropy, Search, Complexity*, volume 16 of *Bolyai Society Mathematical Studies*, chapter 8, Information Theory at the Service of Science, pages 179–207. Springer, 2007.

S. van de Geer. Hellinger-consistency of certain nonparametric maximum likelihood estimators. *The Annals of Statistics*, 21(1):14–44, 1993.

T. van Erven and P. Harremoës. Rényi divergence and majorization. In *IEEE International Symposium on Information Theory (ISIT)*, pages 1335–1339, 2010.

T. van Erven, P. D. Grünwald, and S. de Rooij. Catching up faster by switching sooner: A prequential solution to the AIC-BIC dilemma. *Preprint posted on the math arXiv, arXiv:0807.1005 [math.ST]*, July 2008a.

T. van Erven, P. D. Grünwald, and S. de Rooij. Catching up faster in Bayesian model selection and model averaging. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, pages 417–424. MIT Press, 2008b.

N. K. Vereshchagin and P. M. B. Vitányi. Kolmogorov's structure functions and model selection. *IEEE Transactions on Information Theory*, 50 (12):3265–3290, 2004.

P. Vitányi. Algorithmic statistics and Kolmogorov's structure functions. In P. D. Grünwald, I. J. Myung, and M. A. Pitt, editors, *Advances in Minimum Description Length: Theory and Applications*, chapter 6, pages 151–174. MIT Press, 2005.

P. Volf and F. Willems. Switching between two universal source coding algorithms. In *Proceedings of the Data Compression Conference, Snowbird, Utah*, pages 491–500, 1998.

V. Vovk. Derandomizing stochastic prediction strategies. *Machine Learning*, 35:247–282, 1999.

S. Walker. New approaches to Bayesian consistency. *The Annals of Statistics*, 32(5):2028–2043, 2004.

C. S. Wallace and D. M. Boulton. An information measure for classification. *The Computer Journal*, 11(2):185–194, 1968.

C. S. Wallace and P. R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society. Series B*, 49(3):240–265, 1987.

L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2005. Corrected second printing.

L. Wasserman. *All of Nonparametric Statistics*. Springer, 2006.

C. Z. Wei. On predictive least squares principles. *The Annals of Statistics*, 20(1):1–42, March 1992.

Wikipedia entry on *emerald*. http://en.wikipedia.org/w/index.php?title=emerald&oldid=348923501, March 2010.

H. Wong and B. Clarke. Improvement over Bayes prediction in small samples in the presence of model uncertainty. *The Canadian Journal of Statistics*, 32(3):269–283, 2004.

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

Y. Yang. Mixing strategies for density estimation. *The Annals of Statistics*, 28(1):75–87, 2000.

Y. Yang. Can the strengths of AIC and BIC be shared? *Biometrica*, 92(4): 937–950, 2005.

Y. Yang. Consistency of cross-validation for comparing regression procedures. *The Annals of Statistics*, 35(6):2450–2473, 2007a.

Y. Yang. Prediction/estimation with simple linear models: Is it really that simple? *Econometric Theory*, 23:1–36, 2007b.

Y. Yang. Model selection for nonparametric regression. *Statistica Sinica*, 9:475–499, 1999.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27:1564–1599, 1999.

B. Yu and T. P. Speed. Data compression and histograms. *Probability Theory and Related Fields*, 92:195–229, 1992.

T. Zhang. From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5): 2180–2210, 2006.

Summary

According to the minimum description length (MDL) principle, data compression should be taken as the main goal of statistical inference. From this perspective probability distributions and models are viewed as strategies for data compression, which stands in sharp contrast to making assumptions about an underlying "true" distribution generating the data, as is standard in the traditional frequentist approach to statistics. Strategies are either good or bad, and certainly one should not expect bad models to magically lead to good inference. But, unlike assumptions, strategies can never be true or false. Therefore, if the MDL premise of making data compression a fundamental notion can hold its ground, it promises a robust kind of statistics, which does not break down when standard, but hard to verify, assumptions are not completely satisfied.

This makes it worthwhile to put data compression to the test, and see whether it really makes sense as a foundation for statistics. A natural starting point are cases where standard MDL methods show suboptimal performance in a traditional frequentist analysis. This thesis analyses two such cases. The first case, studied in Part I of the thesis, deals with switching between prediction strategies. The second case, described in Part II, deals with a modification of the standard MDL estimator proposed in the literature, which goes against its data compression motivation.

Part I In Chapter 2 of Part I we investigate cases in which the standard MDL method for model selection leads to suboptimal predictions

of future data. It is found that this may be explained by the fact that there exist shorter descriptions of the data than the descriptions used by standard MDL. Based on this insight, we modify the MDL estimator such that it can use these shorter descriptions and show that this resolves the problem. Thus the standard MDL method fails, but data compression still makes sense and actually leads to the solution of the problem. As a by-product, our investigations shed new light on an old discussion in statistics about whether one should use an AIC-type method or a BIC-type method for model selection.

The shorter descriptions found in Chapter 2 are based on combinations of the models that use a different model for different parts of the data. In Chapter 3 a new method is introduced that automatically determines the optimal bias towards splitting the data into more parts. In Chapter 4 we discuss whether the parts should be modelled independently, or as part of the rest of the data. A new method is introduced to deal with the first case, which is appropriate, for example, for certain time series data.

Part II In Part II we also study the quality of predictions based on the MDL estimator, and investigate under which conditions they converge to the best possible predictions. In order to prove a very general convergence result, previous authors have proposed to modify the standard estimator in a way that, contrary to the data compression philosophy, *increases* the description length of the data. Chapter 5 provides a preliminary discussion of whether this modification is really necessary. Examples are provided showing that no general convergence result can be obtained if the modification is simply omitted, but then it is also shown that in certain common settings no modification is necessary. Although in this case no final verdict is reached on the appropriateness of data compression as a fundamental principle, a technical refinement of existing methods is introduced and the results suggest interesting directions for future study.

The cases identified in Chapter 5 which do not require any modification of the standard MDL estimator, are characterized using a measure of dissimilarity between probability distributions called Rényi divergence. Although Rényi divergence has been around for almost fifty years and appears in many proofs, there exists no overview of its technical properties. Chapter 6 remedies this situation by formally proving the basic properties of Rényi divergence.

Samenvatting

Het minimum description length (MDL) principe schrijft voor dat een dataset het best wordt samengevat door zijn kortst-mogelijke beschrijving. Dit leidt tot een interpretatie van kansverdelingen en statistische modellen als strategieën om experimentele data zo kort mogelijk te beschrijven, en is een radicaal andere insteek dan de traditionele frequentistische benadering van de statistiek, waarin het gebruikelijk is om aannames te doen over een "ware" kansverdeling volgens welke de data gegenereerd zouden worden. Strategieën kunnen slim of dom zijn, en men kan zeker niet verwachten dat slechte modellen op magische wijze tot goede statistische inzichten zullen leiden. Maar in tegenstelling tot aannames, zijn strategieën nooit waar of onwaar. De MDL-insteek, waarin het zo kort mogelijk beschrijven van de data tot het hoofddoel van de statistische analyse wordt verheven, belooft daarom een robuust soort statistiek, die niet direct faalt wanneer standaard, maar moeilijk te controleren, aannames niet volledig opgaan.

Dit maakt het de moeite waard om de kwaliteit van korte beschrijvingen van de data onder de loep te nemen, en te onderzoeken of ze een zinnig fundament voor de statistiek kunnen vormen. Een voor de hand liggend uitgangspunt daarvoor vormen gevallen waarin de standaard MDL-methodes suboptimaal presteren onder een traditionele frequentistische analyse. In dit proefschrift worden twee van dat soort gevallen onderzocht. Het eerste geval, bestudeerd in Deel I, betreft het wisselen tussen voorspelstrategieën. Het tweede geval, dat beschreven wordt in Deel II, gaat over een in de literatuur voorgestelde aanpassing van de standaard MDL-schatter die tegen het MDL-principe indruist.

Deel I In Hoofdstuk 2 van Deel I onderzoeken we gevallen waarin de standaard MDL-methode voor modelselectie tot suboptimale voorspellingen van toekomstige data leidt. Een verklaring wordt gevonden in het feit dat er kortere beschrijvingen van de data mogelijk zijn dan de beschrijvingen die standaard MDL gebruikt. Gebruikmakend van dit inzicht kunnen we de MDL-methode zodanig aanpassen dat hij de kortere beschrijvingen kan gebruiken, en we laten zien dat dit het probleem oplost. Ook al werkt de standaard MDL-methode niet optimaal, toch blijkt het zoeken naar zo kort mogelijke beschrijvingen dus te werken, en leidt het in dit geval zelfs naar de oplossing van het probleem. Een bijproduct van onze analyse is dat hij nieuw inzicht oplevert in een reeds lang lopende discussie in de statistiek over de vraag of men het beste methodes van het AIC- of van het BIC-type kan gebruiken voor modelselectie.

De kortere beschrijvingen uit Hoofdstuk 2 zijn gebaseerd op combinaties van modellen, waarin verschillende modellen worden gebruikt om verschillende delen van de data te beschrijven. In Hoofdstuk 3 wordt een nieuwe methode geïntroduceerd die automatisch bepaalt in hoeveel delen de data het best verdeeld kunnen worden. In Hoofdstuk 4 bespreken we of de verschillende delen het best onafhankelijk van elkaar of juist als onderdeel van de rest van de data beschreven kunnen worden. Er wordt een nieuwe methode geïntroduceerd die geschikt is voor bijvoorbeeld bepaalde tijdsafhankelijke data.

Deel II Net als in Deel I, bestuderen we ook in Deel II hoe goed standaard MDL voorspelt. In dit deel onderzoeken we onder welke voorwaarden de MDL-voorspellingen convergeren naar de best-mogelijke voorspellingen. Om een zeer algemene convergentiestelling te kunnen bewijzen, hebben eerdere auteurs voorgesteld om de standaard MDL-schatter aan te passen, op een manier die leidt tot een *langere* beschrijving van de data en daarmee tegen het MDL-principe ingaat. Hoofdstuk 5 bevat een voorlopige behandeling van de vraag of deze aanpassing werkelijk nodig is. Er worden twee voorbeelden gegeven die laten zien dat een algemene convergentiestelling niet opgaat als de aanpassing van de MDL-schatter achterwege wordt gelaten. Echter, daarna wordt aangetoond dat deze aanpassing in enkele typische gevallen toch niet nodig is. Hoewel er geen eindoordeel wordt geveld over de vraag of het zo kort mogelijk beschrijven van de data altijd tot goede

antwoorden leidt, wordt er wel een technische verfijning van bestaande methodes geïntroduceerd en de gepresenteerde resultaten suggereren interessante richtingen voor nader onderzoek.

De gevallen in Hoofdstuk 5, waarin een aanpassing van de MDL-schatter niet nodig is, worden gekarakteriseerd in termen van een maat voor het verschil tussen kansverdelingen die Rényi-divergentie heet. Hoewel Rényi-divergentie al in de jaren zestig van de vorige eeuw gedefinieerd werd en regelmatig voorkomt in wiskundige bewijzen, bestaat er geen overzicht van zijn technische eigenschappen. Dit gemis wordt aangepakt in Hoofdstuk 6, door op een formele manier alle elementaire eigenschappen van Rényi-divergentie op een rij te zetten.

Curriculum Vitae

Tim van Erven was born twenty-eight years ago in Eindhoven, the Netherlands. Eighteen years later, in 2000, he completed his gymnasium education (grammar school) at the Lorentz-Casimir Lyceum, and then moved to Amsterdam to study artificial intelligence at the University of Amsterdam. During his studies he spent his time with student organisation VIA (as a member of the board in 2004), which supports students of artificial intelligence, business informatics, and computer science at the University of Amsterdam. Nevertheless, both his bachelor's degree (2003) and master's degree (2006) were obtained *cum laude*. The next four years were spent at the Centrum Wiskunde & Informatica (CWI), working as a PhD student under supervision of prof. dr. Peter Grünwald, which resulted in the present thesis. During this time he spent two months visiting prof. Bob Williamson in Canberra, Australia. A preprint of the paper on which Chapter 2 is based, won second prize in the student paper competition of the Risk Analysis section of the American Statistical Association.