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

Tim van Erven Peter Grünwald Steven de Rooij
Centrum Wiskunde & Informatica (CWI)
Science Park 123, P.O. Box 94079
NL-1090 GB Amsterdam, The Netherlands
{Tim.van.Erven,Peter.Grunwald,S.de.Rooij}@cwi.nl

July 7, 2011

Abstract

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. The switch distribution has an efficient implementation. We compare its performance to AIC, BIC and Bayes on a regression problem with simulated data.

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 paper 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 (its computational efficiency is discussed in Section 2.3) 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 $p_1, p_2, \ldots$; 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 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 1.2. Our first main result, Theorem 2 in Section 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 $(\log n)^{2+\alpha}$ for some $\alpha > 0$, 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 5, which is based on a version of the switch distribution that uses only two prediction strategies per considered model, and therefore has a 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 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 8 (Section 8.2), 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.

1.1 Main Application: a Cumulative Risk Version of 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 7. 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 8).

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 paper.

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:

$$p_k(x^n) = \int_{\theta \in \Theta_k} p_{k,\theta}(x^n) w_k(\theta) \, \mathrm{d}\theta. \tag{1}$$

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 $p_{\mathrm{bma}}(x^n) = \sum_k p_k(x^n)\pi(k)$. BMA predicts any new outcome $x_{n+1} \in \mathcal{X}$ outside of the sample x^n according to $p_{\mathrm{bma}}(x_{n+1} \mid x^n)$, which is equal to a combination of the models' predictions in which the models are weighted according to their posterior probability:

$$p_{\mathrm{bma}}(x_{n+1} \mid x^n) = \sum_k p_k(x_{n+1} \mid x^n)\pi(k \mid x^n). \tag{2}$$

We now discuss how the predictions $p_{\mathrm{bma}}(x_{n+1} \mid x^n)$ and $p_k(x_{n+1} \mid x^n)$ may be interpreted as a continuation of predictions on the sample x^n, and how $-\log p_{\mathrm{bma}}(x^n)$ 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 $x^n = x_1, \ldots, x_n$, like for example p_k or p_{bma}. Then the *negative log-likelihood* $-\log p(x^n)$ may 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 $x_1, \ldots, x_n$ 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. Let the $(i+1)$-th outcome be 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 be measured by the *log loss* $-\log p(x_{i+1} \mid x^i)$. Here and in the remainder we let log denote the binary logarithm, such that log loss is measured in *bits*. Summing up the prediction errors, we see that the negative log-likelihood of the sample is equal to the cumulative log loss of the predictions:

$$\sum_{i=1}^{n} -\log p(x_i \mid x^{i-1}) = -\log \prod_{i=1}^{n} p(x_i \mid x^{i-1}) = -\log p(x^n). \tag{3}$$

In particular, $-\log p_{\mathrm{bma}}(x^n)$ and $-\log p_k(x^n)$ 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 $p_{\mathrm{bma}}(x_{n+1} \mid x^n)$, the posterior probability $\pi(k \mid x^n) \propto p_k(x^n)\pi(k)$ may also be interpreted in terms of cumulative loss: 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 cumulative prediction error as measured by log loss. Note that the ratio of posterior probabilities of two models is *exponential* in their difference in cumulative loss!

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 when sequentially predicting x^n. Then prediction using BMA guarantees that the difference between our cumulative loss and the cumulative loss achieved by $\hat{k}$ is in the range $[0, -\log \pi(\hat{k})]$, whatever data x^n are observed. (This follows by (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 paper is that it is often possible to combine the predictions of the models in a way that achieves smaller cumulative loss even

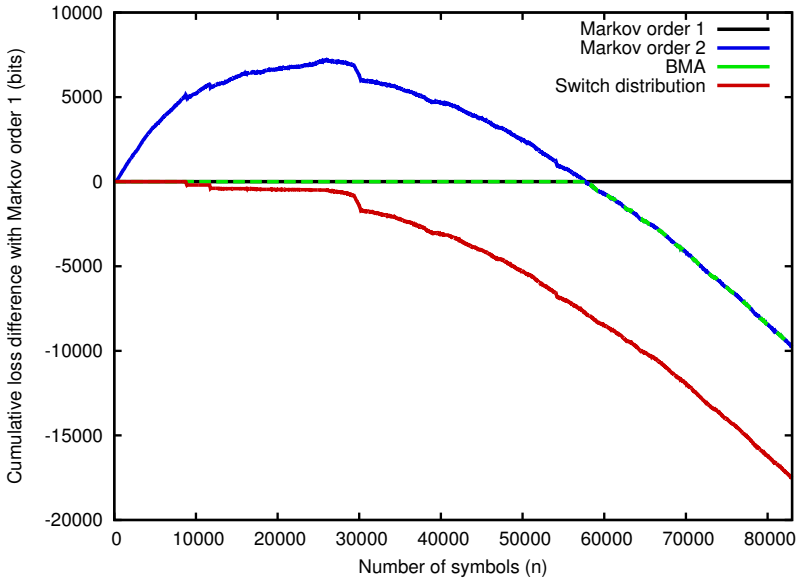


Figure 1: The Catch-up Phenomenon

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. We now give two examples, the first one, based on Figure 1, involving Markov models and the second one based on the normal location family.

The figure compares the cumulative loss of 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. It shows the 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, such that $\mathcal{M}_k$ has $84^k \cdot 83$ parameters. 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.[1]

Note that if the difference in negative log-likelihood 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) *slope* 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, as also shown in the figure, 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 slope, but on the *height* of the graph, and is exponentially concentrated on the model with smallest cumulative loss. 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!

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 1 is intended as a simple illustration of the catch-up phenomenon only. It shows that, if the difference in the number of parameters between $\mathcal{M}_1$ and $\mathcal{M}_2$ is really large, then the catch-up phenomenon can be substantial. Our second example, which highlights the connection to estimation, shows that the phenomenon is still present, though at a smaller scale, if $\mathcal{M}_2$ contains the true distribution, and has just one additional parameter over $\mathcal{M}_1$.

Let $\mathcal{M}_1 = \{\mathcal{N}(0, 1)\}$ correspond to the mean of a Gaussian being zero, and $\mathcal{M}_2 = \{\mathcal{N}(\mu, 1) \mid \mu \in \mathbb{R}\}$

[1] The total cumulative losses for the full book are 603 906, 554 494, 554 495, and 546 698 bits for the first order and second order Markov chains, BMA and the switch distribution respectively.

to it being nonzero. Suppose that $X_1, X_2, \ldots$ are independently distributed according to a distribution $\mathcal{N}(\mu^*, 1) \in \mathcal{M}_2$ with mean $\mu^* \neq 0$ that is close to 0. For convenience, let μ^* be such that $(\mu^*)^2 = 1/n_0$ for some integer n_0. Then, given a sample of size n, the quadratic risk for the maximum likelihood estimator $\hat{\mu}$ in $\mathcal{M}_2$ is $E(\mu^* - \hat{\mu})^2 = 1/n$. And based on $\mathcal{M}_1$, we would always estimate with $\mu = 0$, which has quadratic risk $E(\mu^* - 0)^2 = (\mu^*)^2$. Hence, to get the smallest risk, we should ideally estimate with model $\mathcal{M}_1$ up to sample size n_0, and switch to $\mathcal{M}_2$ for $n > n_0$. The same holds for the expected prediction error if we use the ML estimator for prediction with the logarithmic loss: the expected difference in log loss between predicting with $\hat{\mu}$ and $\mu = 0$ at sample size n is $E[-\log p_{\hat{\mu}}(X_{n+1}) + \log p_{\mu}(X_{n+1})] = \frac{1}{2}(1/n - (\mu^*)^2)$. Nothing essential changes either if we replace estimation using $\hat{\mu}$ by Bayesian prediction based on $\mathcal{M}_2$: let $p_2(X^n)$ be the Bayesian marginal likelihood for Jeffreys' prior or a Normal $\mathcal{N}(0, 1)$ prior. Then the predictive distribution $p_2(X_{n+1} \mid X^n)$ is Gaussian with variance $1 + O(1/n)$, and the expected loss difference between the models becomes $r(n) := \frac{1}{2}(1/n - (\mu^*)^2) + O(1/n^2)$ (see e.g. [Grünwald, 2007, Section 12.3.1, (12.50) and Example 12.8, combined with Section 12.2.2, Lemma 12.2, Part 4 and Example 12.3]). Summing $r(n)$ from 1 to n, it follows that the expected cumulative difference in log loss is $(1/2)\log n - (n/2)(\mu^*)^2 + O(1)$. Ignoring the $O(1)$ term, when $\pi(1) = \pi(2)$, we would (roughly) expect the posterior to give larger mass to $\mathcal{M}_2$ for all $n > n_1$, where n_1 is the smallest n such that $(1/2)\log n < (n/2)(\mu^*)^2$, i.e. when $(\log n)/n < 1/n_0$. Because of the exponential concentration of the posterior on the model with the smallest cumulative loss, BMA will tend to follow p_1 (predicting based on $\mathcal{M}_1$, i.e. with $\mu = 0$) until this sample size, and follow p_2 afterwards. However, the instantaneous risk of p_2 is already smaller than that of p_1 at all sample sizes $n > n_0$. On the interval $[n_0, n_1]$, BMA follows p_2 whereas p_1 is optimal for prediction. These informal calculations can be verified by computer simulation. We encounter the same phenomenon in regression with polynomials (see Section 7).

We claim that the general phenomenon that different models predict better at different sample sizes occurs widely, both in theoretical settings and on real-world data. 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 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 1 shows that, in the Markov example, the switch distribution first predicts roughly like p_1, but switches to p_2 almost immediately after it starts making better predictions.[2] It essentially does this no matter what sequence x^n is actually observed (see Section 8.2 for its performance on the Gaussian location models). 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, 2008b, 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.

1.3 Overview

In Section 2 we define the switch distribution and discuss the computational efficiency of its implementation. 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 3, which also discusses model selection in the sequential prediction setting. A first (minor) result is presented in Section 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 5, apply regardless of whether prediction

[2] 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). 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 order Markov chain.

is based on the Bayesian predictive distribution or on parameter estimation. They are followed by our main consistency result in Section 6, which only applies to Bayesian prediction strategies. Section 7 contains a simulation study of linear regression with polynomials. The discussion in Section 8 puts our work in a broader context and explains how it fits into the existing literature. In particular, Section 8.2 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 in the Appendix.

2 The Switch Distribution

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 $d \in \mathbb{N} = \{1, 2, \ldots\}$. 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 1 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 (regular) conditional probabilities of a distribution for X^∞ with density

$$p(x^n) = p(x_1) \cdot p(x_2 \mid x_1) \cdot \ldots \cdot p(x_n \mid x^{n-1}). \tag{4}$$

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 $n \in \mathbb{N}$ 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 Definition

We start with a given, countable set of prediction strategies $\{p_k \mid k \in \mathcal{A}\}$; see Example 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

$$\mathbb{S} = \Big\{ \big((t_1, k_1), \ldots, (t_m, k_m)\big) \in (\mathbb{N} \times \mathcal{A})^m \mid m \in \mathbb{N}, 1 = t_1 < \ldots < t_m \Big\}. \tag{5}$$

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,

$$q_{\mathbf{s}}(x_{n+1} \mid x^n) = p_{k_j}(x_{n+1} \mid x^n) \qquad \text{for } t_j \leq n+1 < t_{j+1}, \tag{6}$$

with the convention that $t_{m+1} = \infty$. 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.

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

Definition 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

$$p_{\mathrm{sw}}(x^n, \mathbf{s}) := q_{\mathbf{s}}(x^n)\, \pi(\mathbf{s}) \tag{7}$$

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

Hence the marginal switch distribution on n outcomes has density

$$p_{\rm sw}(x^n) = \sum_{\mathbf{s}\in\mathbb{S}} q_{\mathbf{s}}(x^n)\,\pi(\mathbf{s}). \tag{8}$$

By Bayes' theorem, the prior π, conditioned on observed data x^n, induces a posterior distribution $p_{\rm sw}(\mathbf{s} \mid x^n) \propto q_{\mathbf{s}}(x^n)\pi(\mathbf{s})$ 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 $K_{n+1}(\mathbf{s}) = k_j$ for j such that $t_j \le n+1 < t_{j+1}$. Then $K_{n+1}(\mathbf{s})$ is the prediction strategy that is used by $q_{\mathbf{s}}$ to predict the $(n+1)$-th outcome. We can then consider, say, the posterior probability assigned to each prediction strategy upon observing x^n:

$$p_{\rm sw}(K_{n+1} = k \mid x^n) = \left(\sum_{\mathbf{s}:K_{n+1}(\mathbf{s})=k} p_{\rm sw}(x^n,\mathbf{s})\right)/p_{\rm sw}(x^n). \tag{9}$$

This quantity is used to define a model selection criterion based on the switch distribution in Section 3.

2.3 Structure of the Prior and Efficient Computation

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

$$\pi((t_1,k_1),\ldots,(t_m,k_m)) = \mu(m)\left(\prod_{j=1}^{m-1}\kappa_{t_j}(k_j)\tau(Z = t_{j+1} \mid Z > t_j)\right)\lambda_{t_m}(k_m). \tag{10}$$

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. Its conditioning in (10) exploits the prior knowledge that $t_{j+1} > t_j$. For all $n \in \mathbb{N}$, 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 5.1. For computational reasons it is convenient to allow κ_n and 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.

Our consistency and convergence rate theorems impose some further conditions on the prior π. For concreteness, we remark that every prior of the following form is compatible with all our results:

$$\mu(m) = 2^{-m}, \tau(t) = 1/(t(t-1)), \text{ and } \kappa_n \text{ and } \lambda_n \text{ are uniform on their support,} \tag{11}$$

as long as the supports of κ_n and λ_n never shrink with n and are at most of polynomial size in n. For this τ, we have $\tau(Z = t_{j+1} \mid Z > t_j) = (t_{j+1}(t_{j+1}-1))^{-1}/(1 - \sum_{i=2}^{t_j}(i(i-1))^{-1}) = t_j/(t_{j+1}(t_{j+1}-1))$.

Example 1. In the Markov chain example of Figure 1, $p_{\rm sw}$ is instantiated as follows. We set $\mathcal{L} = \mathcal{K} = \mathcal{A} = \{1,2\}$, and define the prior π using (11), with the support of κ_n and λ_n equal to $\mathcal{A}$ for all n. For $k \in \mathcal{A}$, p_k, as used in (6), is defined as the Bayesian marginal likelihood (see (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 paper, the prediction strategies p_k will be based on (parametric) models $\mathcal{M}_k$. They will either be Bayesian predictive distributions as in Example 1, or parameter estimators relative to $\mathcal{M}_k$, as explained in Section 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.

Hidden Markov Model and Efficient Computation In [Van Erven, 2010, Chapter 2] an algorithm is presented that sequentially computes the posterior probabilities $p_{\mathrm{sw}}(K_{i+1} = k \mid x^i)$ for $i = 0, \ldots, n-1$. It requires that $\mu(m)$ is a geometric distribution, like in (11). Let $s_i = |\bigcup_{j=1}^{i} \mathrm{support}(\lambda_j)| + |\bigcup_{j=1}^{i} \mathrm{support}(\kappa_j)|$ denote the number of strategies considered by time i. Then the total running time of the algorithm is $O(\sum_{i=1}^{n} s_i)$, which is linear in the number of outcomes n and the sizes of the supports. For example, if $\mathrm{support}(\kappa_i) = \mathrm{support}(\lambda_i) = \mathcal{A}$ for all i, 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 i, see Section 5.3, Example 2.

The algorithm is closely related to algorithms for tracking the best expert from the universal prediction literature. It may also be interpreted as an instance of the *forward algorithm* for a hidden Markov model with hidden variables $K_1, K_2, \ldots$ This perspective is taken by Koolen and de Rooij [2008a], who also verify that this hidden Markov model corresponds exactly to the switch distribution as defined above.

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 cumulative log loss 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 (10). We therefore find that

$$p_{\mathrm{sw}}(x^n) \geq \sum_{\mathbf{s} \in \{((1,k)) \mid k \in \mathcal{L}\}} \pi(\mathbf{s}) q_{\mathbf{s}}(x^n) = \mu(1) \sum_{k \in \mathcal{L}} \lambda_1(k) p_k(x^n) = \mu(1) p_{\mathrm{bma}}(x^n)$$

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 1 illustrates, the switch distribution may achieve substantially smaller cumulative loss than p_{bma}. This is also seen in the simulation study in Section 7.

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.

3.1 Stage 1: Models and Associated Prediction Strategies

We define a *model* $\mathcal{M}$ as a set of prediction strategies. With each model, we associate a single "meta" prediction strategy; the models themselves are *only used in terms of these meta strategies:* our results about predictive performance in Sections 4 and 5 apply regardless of how these meta strategies are defined; for our consistency result there are some restrictions that are explained in Section 6. For example, a prediction strategy for a parametric model $\mathcal{M} = \{p_\theta \mid \theta \in \Theta \subseteq \mathbb{R}^d\}$ may be defined in terms of a *parameter estimator* $\hat{\theta} : \mathcal{X}^* \to \Theta$. The next outcome is subsequently predicted using $p(x_{n+1} \mid x^n) = p_{\hat{\theta}(x^n)}(x_{n+1} \mid x^n)$. Recall that by (4) this also defines a joint density $p(x^n) = p(x_1 \mid x^0) \cdot \ldots \cdot p(x_n \mid x^{n-1})$. A second important example is to take the Bayesian approach: given a prior density w on Θ, define the marginal likelihood by

$$p(x^n) = \int_{\theta \in \Theta} p_\theta(x^n) w(\theta) \, \mathrm{d}\theta. \tag{12}$$

We obtain a prediction strategy by applying (4) in the other direction: $p(x_{n+1} \mid x^n) = p(x^{n+1})/p(x^n)$. 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].

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 general, the number of parameters in $\mathcal{M}_k$ does not need to be a straightforward function of 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.

Model selection can either be used to obtain adaptive estimators, or to determine which model (if any) contains the true distribution. In the latter case the model selection criterion needs to be consistent, see Section 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). 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 7).

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 $\mathbb{S}$ rather than simply on the models. In prediction, the models are therefore averaged by marginalising the posterior using the random variable K_{n+1} from (9):

$$p_{\mathrm{sw}}(x_{n+1} \mid x^n) = \sum_{k \in \mathcal{A}} p_k(x_{n+1} \mid x^n) p_{\mathrm{sw}}(K_{n+1} = k \mid x^n). \tag{13}$$

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 5). Hence we define the following model selection criterion for the switch distribution, which selects a model index from $\mathcal{L}$ only:

$$\delta_{\mathrm{sw}}(x^n) = \arg\max_{k \in \mathcal{L}} p_{\mathrm{sw}}(K_{n+1} = k \mid x^n). \tag{14}$$

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 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.

4.1 Model Classes

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. To test the predictions of the switch distribution, we will want to assume that X^∞ is distributed according to a distribution p^* from a class $\mathcal{M}^*$ that is not necessarily equal to $\mathcal{M}$.[3] 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 Section 4.4. Our strongest risk convergence results however, presented in Section 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*.

4.2 Risk

The Kullback-Leibler (KL) risk of an estimator p is

$$r(p^*, p, n) = \underset{X^{n-1} \sim p^*}{E} \left[D\big(p^*(X_n \mid X^{n-1}) \| p(X_n \mid X^{n-1})\big) \right], \tag{15}$$

where p^* is the true distribution and $D(p\|q) = E_{Y\sim p}\left[\log \frac{p(Y)}{q(Y)}\right]$ is the KL divergence (which is nonnegative). 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, 1998b] or [Grünwald, 2007, Chapter 15]). This implies the following proposition, which underlies all our convergence rate results:

Proposition 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 $X^\infty \sim p^*$ to be i.i.d.

Remark 1. As we observed in Section 2.4, the switch distribution dominates Bayesian model averaging by a factor $\mu(1)$, so by Proposition 1

$$R(p^*, p_{\mathrm{sw}}, n) \leq R(p^*, p_{\mathrm{bma}}, n) - \log \mu(1),$$

for any p^* whatsoever.

4.3 Minimax Risk Convergence

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:

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

[3] In p^* and $\mathcal{M}^*$, the star is simply part of the name, not the Kleene star operator.

where the infimum is over all possible estimators, as defined in Section 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 [1998b], Yang and Barron [1999] and Poland and Hutter [2005].

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

Definition 2. For any two nonnegative functions $g, h : \mathbb{N} \to [0, \infty]$, 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 paper: 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 8.2 for further discussion of the relationships between these two convergence notions.

4.4 The Parametric Case

Let $\mathcal{M}_k = \{p_\theta \mid \theta \in \Theta_k\}$ be a d-dimensional parametric family, and let p_k be a corresponding Bayesian prediction strategy, defined as in (12). Then Clarke and Barron [1990] show that, under suitable regularity conditions, which include a compactness condition on Θ_k, the cumulative risk of p_k satisfies

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

uniformly for all $p^* \in \mathcal{M}_k$. They also show [Clarke and Barron, 1994] that the minimax cumulative risk relative to the model class $\mathcal{M}^* := \mathcal{M}_k$ satisfies $R_{\mathrm{mm}}(n) = (d/2) \log n + O(1)$. 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 1 implies that p_{bma} also achieves the minimax cumulative risk up to factor 1. Subsequently, Remark 1 implies that the switch distribution also achieves the minimax cumulative risk up to factor 1.

Actually, our experiments for the parametric case in Section 7 suggest that the cumulative risk of the switch distribution may be smaller than that of Bayesian model averaging by a constant, but we have no general theorems to substantiate this.

5 Two Cumulative Risk Bounds

In the nonparametric case, where p^* is in none of the considered models, the minimax optimal cumulative risk grows more quickly than in the parametric case. As a result the cumulative risk of p_{bma} may not be minimax optimal anymore [Rissanen et al., 1992]. In contrast, in this section we establish minimax optimality of the switch distribution.

To present our risk bounds for nonparametric adaptive estimation based on the switch distribution, we first need to introduce the notion of "frozen" prediction strategies, which keep issuing the same prediction even as they are conditioned on more and more data. These 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, our strongest cumulative risk result, is presented in Section 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 5 in Section 5.4.

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., 2008] 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 3 (Standard i.i.d.)**.** We call a distribution p^* for $X^\infty = X_1, X_2, \ldots$ "standard i.i.d." 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 i.i.d." if all $p^* \in \mathcal{M}^*$ are standard i.i.d. For any two standard i.i.d. 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, 1998a, 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 on appropriate intervals. This is achieved by "freezing" the issued predictions as follows.

Definition 4 (Frozen Strategies)**.** Let p_k be a prediction strategy and t a positive integer. Then p_k *frozen at time* t is a new prediction strategy defined by $p_{k \circ t}(x_{n+1} \mid x^n) = p_k(x_{n+1} \mid x^{\min\{t-1, n\}})$. Similarly, for a (finite or infinite) increasing sequence of positive integers $\mathbf{t} = t_1, t_2, \ldots$ with $t_1 = 1$, the strategy p_k *frozen at times* $\mathbf{t}$ is $p_{k \circ \mathbf{t}}(x_{n+1} \mid x^n) = p_k(x_{n+1} \mid x^{t_j - 1})$ when $t_j \leq n+1 < t_{j+1}$, with the convention that $t_{m+1} = \infty$ if $\mathbf{t} = t_1, \ldots, t_m$ is finite.

Note that, when the data are i.i.d., any prediction strategy p_k that is frozen at time t keeps issuing the same predictions for all $n+1 \geq t$ and consequently has a fixed risk. Similarly, a prediction strategy p_k frozen at times $\mathbf{t} = t_1, t_2, \ldots$ has the same fixed risk between any two freezing times t_j and t_{j+1}.

5.2 Oracles, Fast and Slow Switch Distribution

Let $\{p_k \mid k \in \mathcal{L}\}$ be a countable set of prediction strategies, each associated with a corresponding model $\mathcal{M}_k$. We define two advanced versions of the switch distribution, in which $\mathcal{K}$ contains frozen versions of these prediction strategies, and $\mathcal{A} = \mathcal{L} \cup \mathcal{K}$: in the *slow switch distribution*

$$\mathcal{K} = \{k \circ t \mid k \in \mathcal{L}, t \in \mathbb{N}\}; \tag{16}$$

and in the *fast switch distribution*

$$\mathcal{K} = \{k \circ \mathbf{t} \mid k \in \mathcal{L}\} \tag{17}$$

for a fixed increasing sequence $\mathbf{t} = t_1, t_2, \ldots$ that may be chosen by the practitioner. The basic switch distribution corresponds to the special case of the fast switch distribution with $\mathbf{t} = 1, 2, 3, \ldots$ (such that freezing has no effect), but Theorem 5 below will require the freezing times to be exponentially increasing.

We will bound the cumulative risk of these switch distributions by that of an *oracle* [Donoho and Johnstone, 1994] that selects a prediction strategy based on the data and knowledge of the true distribution p^*. Model selection criteria are examples of oracles, but because they know p^* oracles are more powerful. In this paper, we adopt a broad definition that gives the oracle full access to p^*:

Definition 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 with index $\omega(p^*, x^n)$. For $\mathcal{S}_1, \mathcal{S}_2, \ldots$ subsets of $\mathcal{A}$, we say that ω is an oracle *relative to* $\mathcal{S}_1, \mathcal{S}_2, \ldots$ if $\omega(p^*, x^n) \in \mathcal{S}_{n+1}$ for all $p^* \in \mathcal{M}^*$, $n \geq 0$, $x^n \in \mathcal{X}^n$. 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.

Our theorems will apply to oracles relative to sets $\mathcal{L}_1 \subseteq \mathcal{L}_2 \subseteq \ldots \subseteq \mathcal{L}$, where the size of $\mathcal{L}_n$ grows at most polynomially in n. To formulate the condition that enforces this requirement, define

$$\begin{aligned} \mathcal{K}_n &= \{k \circ t \mid k \in \mathcal{L}_n, t \in \{1, \ldots, n\}\} \subset \mathcal{K} && \text{(slow switch distribution)} \\ \mathcal{K}_n &= \{k \circ \mathbf{t} \mid k \in \mathcal{L}_n\} \subset \mathcal{K} && \text{(fast switch distribution),} \end{aligned} \tag{18}$$

analogously to (16) and (17). Then we impose the following condition on the prior:

Condition 1. The prior π of the switch distribution is defined as in (10) and satisfies

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

where $\mathcal{K}_n$ is as defined in (18).

This condition implies that the tails of the distributions τ and κ_n are of polynomial thickness, and that $\mathcal{K}_n$ is at most polynomially large in n. As $|\mathcal{K}_n| = n|\mathcal{L}_n|$ for the slow switch distribution, and $|\mathcal{K}_n| = |\mathcal{L}_n|$ for the fast switch distribution, this implies that $\mathcal{L}_n$ is also at most polynomially large. Thus, the number of models we allow an oracle to consider is at most polynomial in n.

Example 2. Suppose that $\mathcal{L} = \mathbb{N}$. 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 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 Condition 1, having infinite supports for λ_n and κ_n leads to computational issues. These can be addressed by reducing the supports of λ_n and κ_n to $\mathcal{L}_n$ and $\mathcal{K}_n$ (or suitably small supersets), respectively, and using the sample size dependent prior suggested in (11). The resulting running time for the algorithm on data $x_1, \ldots, x_n$ will then be of order $\sum_{i=1}^{n} (|\mathcal{K}_i| + |\mathcal{L}_i|)$, which is $O(n^{2+a})$ and $O(n^{1+a})$ for the slow and fast switch distributions, respectively.

The difference in running time in Example 2 motivates the adjectives "slow" and "fast" for the two switch distributions. As will be seen in the next two sections, the smaller running time of the fast switch distribution does come at a price: whereas, under weak conditions, the slow switch distribution achieves the minimax cumulative risk up to factor one (which is optimal), 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.

5.3 Cumulative Risk Bound for the Slow Switch Distribution

The cumulative risk of the slow switch distribution asymptotically grows at the same rate as that of any oracle, provided that the cumulative risk of that oracle is not too small:

Theorem 2 (Cumulative Risk for Slow Switch Distribution)**.** *Fix $\mathcal{L}_1 \subseteq \mathcal{L}_2 \subseteq \cdots \subseteq \mathcal{L}$, let $\mathcal{M}^*$ be standard i.i.d. and choose a prior that satisfies Condition 1. Then, for any oracle ω relative to $\mathcal{L}_1, \mathcal{L}_2, \ldots$ that satisfies*

$$\frac{(\log n)^{2+\alpha}}{R_{\mathrm{m}}(p_\omega, n)} \to 0 \tag{19}$$

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

$$R_{\mathrm{m}}(p_{\mathrm{sw}}, n) \preceq R_{\mathrm{m}}(p_\omega, n). \tag{20}$$

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 (19) 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 4 below) $R_{\mathrm{m}}(p_\omega, n) \geq R_{\mathrm{mm}}(n) \succeq n r_{\mathrm{mm}}(n) \succeq n^{1-\gamma}$, and (19) holds. Hence Theorem 2 implies the following:

Corollary 3. *Suppose $\mathcal{M}^*$ is a standard i.i.d. 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 $\delta : \mathcal{X}^n \to \mathcal{L}_{n+1}$, which selects only prediction strategies from $\mathcal{L}_{n+1}$, the worst-case cumulative risk of the slow switch distribution (with a prior that satisfies Condition 1) grows no faster than the worst-case cumulative risk of δ. That is,*

$$R_{\mathrm{m}}(p_{\mathrm{sw}}, n) \preceq R_{\mathrm{m}}(p_\delta, n), \tag{21}$$

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 "nonparametric", 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.

Example 3. 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. For each J, they construct a fixed smooth prior density w_J on the canonical parameters of $\mathcal{M}_J$. Now suppose that the true density p^* belongs to the class of α-smooth functions $C^\alpha[0, 1]$. Then it follows from Theorem 5.2 of Ghosal et al. that the Bayesian prediction strategy with prior w_J of dimension $J = J_{n,\alpha} = \lfloor n^{1/(2\alpha+1)} \rfloor$ achieves the minimax rate of convergence $n^{-\alpha/(2\alpha+1)}$ in Hellinger risk. 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 minimax rate $n^{-2\alpha/(2\alpha+1)}$ in KL risk.

To obtain an adaptive procedure, they consider various priors on α. In particular, they show that by putting a discrete prior on the set of rational-valued smoothnesses $\alpha \in \mathbb{Q}^+$, the optimal rate is achieved up to a logarithmic factor, which they believe "is not a defect of [their] proof, but connected to this prior." In this case the same prior would 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, which makes Bayes prefer overly simple models (with too small $J_{n,\alpha}$ or, equivalently, too large α). Indeed, Ghosal et al. show that for an alternative, sample size-dependent prior on α that puts less mass on small models, the extra logarithmic factor is avoided.

In terms of cumulative risk, the logarithmic factor is also avoided by the switch distribution, even without the use of any sample size-dependent priors: for $J \in \mathcal{L} = \mathbb{N}$ define Bayesian prediction strategies p_J based on the same priors w_J as Ghosal et al., and take $\lambda_n(J) = 1/J(J+1)$ and $\kappa_n(J \circ t) = 1/(J(J+1)\, t(t+1))$ independent of n as in Example 2. In addition any suitable μ and τ such that Condition 1 holds may be chosen. Note that neither $\mathcal{L}$ nor the prior depend on α, which therefore does not have to be known in advance. Now Theorem 2 may be applied with the oracle $\omega(p^*, x^n) = \lfloor n^{1/(2\alpha+1)} \rfloor$ and $\mathcal{L}_n = \{1, \ldots, \lfloor n^{1/(2\alpha+1)} \rfloor\}$, showing that the switch distribution adaptively achieves the minimax cumulative rate, which is $n^{1/(2\alpha+1)}$ by Proposition 4 below.

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 1 implies that the model selection criterion δ mentioned in Corollary 6 must output a model with index in a set that grows at most 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 8.4.

3. Since ω in Theorem 2 is an oracle with access to p^*, it may in some cases perform substantially better than any (known) model selection criterion that does not know p^*. Thus, in principle, Theorem 2 is much stronger than Corollary 3.

4. The theorem is asymptotic, but by going through the steps of the proof with $\alpha = 0$ and keeping track of constants, one can also show that $R_{\mathrm{m}}(p_{\mathrm{sw}}, n) \leq 2R_{\mathrm{m}}(p_\omega, n) + c_1(\log n)^2 + c_2$, 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.

5. The theorem is interesting only if $R_{\mathrm{m}}(p_\delta, n)$ is finite, which implies that $R_{\mathrm{m}}(p_{\mathrm{sw}}, n)$ is finite, and therefore $r_{\mathrm{mm}}(i) \leq R_{\mathrm{mm}}(n) \leq R_{\mathrm{m}}(p_{\mathrm{sw}}, n)$ should be finite as well, for all $i \leq n$. Note, however, that when p^* is standard i.i.d., finiteness of $r_{\mathrm{mm}}(1)$ implies finiteness of $r_{\mathrm{mm}}(i)$ for all $i \geq 1$ and hence finiteness of $R_{\mathrm{mm}}(n) \leq \sum_{i=1}^n r_{\mathrm{mm}}(i)$.

5.4 Cumulative Risk Bound for the 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. 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

$$c_1 h_0(n) \preceq n^\gamma r_{\mathrm{mm}}(n) \preceq c_2 h_0(n). \tag{22}$$

As can be seen by inspecting the proof of Theorem 5 below, this condition implies that $(\log n)^{2+\alpha}/R_{\mathrm{mm}}(n) \to 0$ for any α and is therefore stronger than what is required for Corollary 3. Yet, it is still weak enough to be satisfied by all model classes that are usually called nonparametric. 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 5 below.

If Condition 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 Appendix A.3:

Proposition 4. *Suppose that $\mathcal{M}^*$ is a standard i.i.d. 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 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 Appendix A.4 we prove the following theorem:

Theorem 5 (Cumulative Risk for Fast Switch Distribution)**.** *Fix $\mathcal{L}_1 \subseteq \mathcal{L}_2 \subseteq \cdots \subseteq \mathcal{L}$, let $\mathcal{M}^*$ be a standard i.i.d. model class that satisfies Condition 2 with constants c_1, c_2 and γ, and choose a prior that satisfies Condition 1. Suppose there exists an oracle ω relative to $\mathcal{L}_1, \mathcal{L}_2, \ldots$ that achieves the minimax risk up to a nondecreasing function $f : \mathbb{N} \to [1, \infty)$, 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. Then for any infinite increasing sequence of freezing times $\mathbf{t} = t_1, t_2, \ldots$ with $t_1 = 1$ and $t_j \geq a \exp(bj)$ for positive constants a and b, the corresponding fast switch distribution achieves the minimax cumulative risk up to factor $cf(n)$ for a constant c. Specifically,*

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

with c given by

$$c = \left(\frac{c_2}{c_1}\right)^2 \cdot \frac{1}{1-\gamma} \sup_{j \geq 1} \left(\frac{t_{j+1} - 1}{t_j}\right)^\gamma. \tag{23}$$

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 (23) is bounded by $(1+\epsilon)^\gamma$. Analogously to Corollary 3, Theorem 5 implies the following:

Corollary 6. *Suppose $\mathcal{M}^*$ is a standard i.i.d. model class that satisfies Condition 2. Let the fast switch distribution be as in Theorem 5. If there exists any model selection criterion $\delta\colon \mathcal{X}^n \to \mathcal{L}_{n+1}$ 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* (23)*, 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 always achieves the minimax cumulative risk, albeit only up to a factor c', which may be larger than 1. Remarks analogous to remarks 1–5 below Corollary 3 apply to Corollary 6 as well.

6 Consistency

In Section 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 7 (Consistency of the Switch Distribution)**.** *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* (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 $B^k_n = \{((t_1, k_1), \ldots, (t_m, k_m)) \in \mathbb{S} \mid t_m \leq n+1, k_m = k\}$ 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*

$$\sum_{\mathbf{s} \in B^k_{n_k}} \pi(\mathbf{s}) q_{\mathbf{s}}(X^{n_k}) > 0 \qquad (p_k\text{-a.s.}) \tag{24}$$

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 on K_{n+1} satisfies

$$p_{\mathrm{sw}}(K_{n+1} = k^* \mid X^n) \stackrel{n\to\infty}{\longrightarrow} 1 \qquad \text{with } p_{\theta^*}\text{-probability } 1, \tag{25}$$

which implies consistency of δ_{sw} as defined in (14)*.*

For $k \in \mathcal{L}$ such that $\lambda_1(k)$ is positive, (24) 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 equals $\mathcal{L}$. 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 7 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 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 and Markov order estimation (as in the introductory example). In 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.

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 proven to achieve minimax risk. However, these estimators are often not Bayesian, so that the requirements of Theorem 7 are not satisfied.

There are two ways to bridge this gap: first, one might show that the risk of a Bayesian estimator is so close to the risk of an estimator that is known to achieve minimax risk, that it must achieve the minimax risk itself. This can be done for Gaussian regression with random design, where Bayesian prediction strategies based on Jeffreys' prior are sufficiently similar to least-squares estimators. See [Van Erven, 2010, Chapter 2] for details. The experimental set-up in the next section is a special case. In similar fashion, it is possible to establish both minimax cumulative risk and 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]).

Secondly, one might extend Theorem 7 to non-Bayesian estimators. An initial such generalisation is provided by [Van Erven et al., 2008].

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. We compare six methods: $\mathcal{C} =$ {Fast switch, Slow switch, Basic switch, Bayes, AIC, BIC}.

The set-up is as follows: let $(X_1, Y_1), (X_2, Y_2), \ldots$ be independent, identically distributed pairs of random variables, with X_i sampled uniformly at random from $[-1, 1]$ and

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

for some regression function f^* and normally distributed noise ξ_i with mean zero and known variance $\sigma^2 = 1$. The regression function is approximated by polynomials

$$\mathcal{F}_k = \left\{\theta_k x^k + \ldots + \theta_1 x + \theta_0 \mid (\theta_0, \ldots, \theta_k) \in \mathbb{R}^{k+1}\right\}$$

of degree $k \in \{0, \ldots, K\}$. For every polynomial f, let $p_f(Y_i \mid X_i)$ be the density of the normal distribution with mean $f(X_i)$ and variance σ^2, and let $u(X_i) = 1/2$ be the uniform density on $[-1, 1]$. Then define the probabilistic models

$$\mathcal{M}_k = \left\{p_f(Y_i \mid X_i)u(X_i) \mid f \in \mathcal{F}_k\right\},$$

for which the maximum likelihood estimator equals the least squares estimator in $\mathcal{F}_k$. We note that since $u(X_i)$ does not depend on f or k, it does not influence any of the procedures we are comparing, so its choice is unimportant.[4]

With each model $\mathcal{M}_k$, we associate a Bayesian prediction strategy p_k based on Jeffreys' prior, which makes the posterior distribution on the parameters of $\mathcal{M}_k$ multivariate normal with mean $(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'\mathbf{y}$ and covariance matrix $\sigma^2(\mathbf{X}_k'\mathbf{X}_k)^{-1}$, where $\mathbf{X}_k$ is the $n\times(k+1)$ design matrix with entries $(\mathbf{X}_k)_{ij} = X_i^{j-1}$ and $\mathbf{y} = (Y_1,\ldots,Y_n)'$ [Box and Tiao, 1973]. The resulting predictions are $p_k(X_{n+1},Y_{n+1} \mid (X,Y)^n) = p_k(Y_{n+1} \mid X_{n+1},(X,Y)^n)\cdot u(X_{n+1})$, where $p_k(Y_{n+1} \mid X_{n+1},(X,Y)^n)$ is a normal distribution with mean $(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'\mathbf{y}$ and variance $\sigma^2(1+X_{n+1}(\mathbf{X}_k'\mathbf{X}_k)^{-1}X_{n+1})$ [Grünwald, 2007]. Although Jeffreys' prior is improper, this predictive distribution is almost surely well-defined when the number of observations is at least $k+1$. To ensure the predictions for all models are well-defined, we therefore evaluate cumulative risk for $n \geq K+1$.

With each method in $\mathcal{C}$, we associate a model selection criterion and an estimator. The basic, slow and fast switch distributions are defined as in Sections 2.2 and 5.2; the associated model selection criteria are given by (14). We use $\mathcal{L}_n = \mathcal{L} = \{0,1,\ldots,K\}$ and in case of the fast switch distribution choose freezing times $t_j = \max\{j, \lfloor 1.1^{j-1} \rfloor\}$ for $j = 1,2,\ldots$ The priors are chosen as in (11), where the supports of λ_n and κ_n are $\mathcal{L}_n$ and $\mathcal{K}_n$ (see (18)), 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 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 1 TEST(f^*, N, K, R)

for $r = 1,\ldots,R$ **do**
  for $n = K+1,\ldots,N$ **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 empirical risk $\log_2 \left(\frac{\varphi(y_n - f^*(x_n))}{p\left(y_n \mid x^n, y^{n-1}\right)} \right)$, 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 $K = 35$.

[4] In fact, it is equivalent to drop $u(X_i)$ altogether and work with the conditional densities $p_f(Y_i \mid X_i)$ only, which makes the procedures discriminative instead of generative. It is also straightforward to extend our theoretical results to cover this case.

(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 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(c) 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(e) 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 correction for AIC, such as in AIC_c [Burnham and Anderson, 2002].

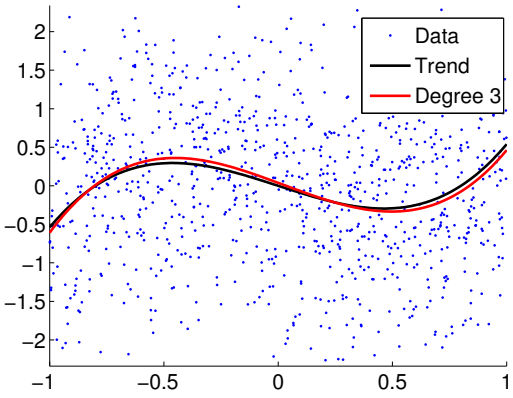
In this parametric experiment, eventually all consistent methods select $\mathcal{M}_3$, so their instantaneous risks converge to the instantaneous risk of p_3. Thus, the difference in cumulative risk for these methods will converge to a constant. In fact, by $n = 1000$ the lines for each method already 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, 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. Again, 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.

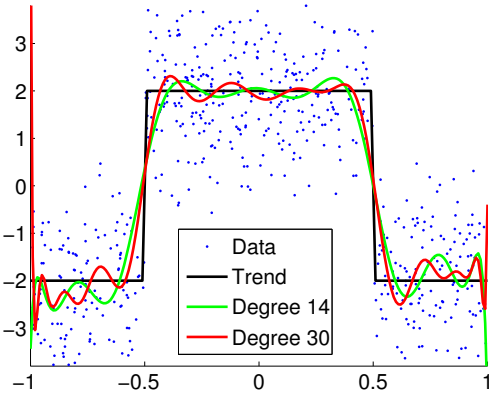
Interpretation The experiments confirm the theoretical results of this paper: (1) all considered methods except AIC are consistent, (2) BIC and Bayes suffer from the catch-up phenomenon and as such issue inferior predictions. The predictive performance of the switch distribution, at least in its fast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

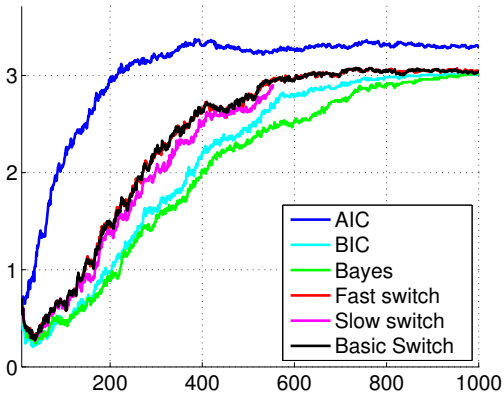
This is clearly an issue that deserves careful consideration in practice if the cumulative risk is very small. Whether or not it is small depends very much on the setting; recall that in the Markov chain example in the introduction a single switch yielded a reduction in cumulative loss of about 7 000 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



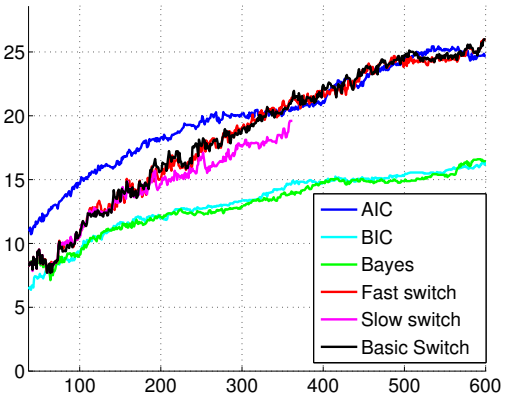
(a) Typical data



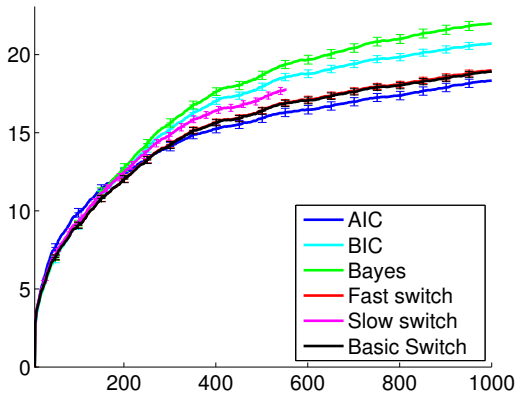
(b) Typical data



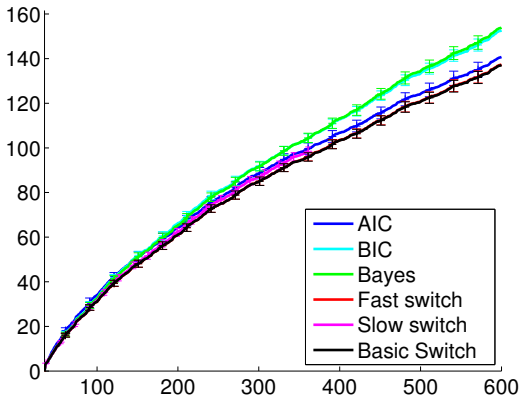
(c) Average degree selected



(d) Average degree selected



(e) Estimated cumulative risk



(f) Estimated cumulative risk

Figure 2: Sequential polynomial regression results

(see Remark 1).

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.

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 statisticians, 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, 1998b, Yang and Barron, 1999, 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 cannot solve the problem in general, as is discussed in Section 8.2.

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. The main difference between our approach and Yang's is that we use a single paradigm rather than a combination 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.

8.2 Cumulative vs Instantaneous Risk

In the parametric case, based on Theorem 7 and the discussion in Section 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. To see what happens, reconsider the normal location family example of Section 1.2. A procedure that achieves minimax instantaneous risk would have to switch from $\mathcal{M}_1$ to $\mathcal{M}_2$ at a sample size sufficiently close to n_0. The switch distribution prior divides its mass over switches at each sample size n. The total prior mass for sample sizes close enough to n_0 is apparently too small to achieve the 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 5.3 and 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

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

We have the following result, proved in Appendix A.3:

Proposition 8. *Suppose $\mathcal{M}^*$ is a standard i.i.d. model class that satisfies Condition 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 applies, p_{sw} also achieves the minimax instantaneous risk, and hence also resolves this version of the AIC-BIC dilemma.

8.3 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".

8.4 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 $\mathcal{K}$. The basic switch distribution corresponds to $\mathcal{K} = \mathcal{L}$, which works well in practice but invalidates the proofs of Theorems 2 and 5. 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 can be shown for the basic switch distribution? 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 1, sequentially predicting using the prediction strategy $p_{\delta_{\text{sw}}(x^n)}$ for the model with index $\delta_{\text{sw}}(x^n)$, 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 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.

Acknowledgements

We would like to thank an anonymous referee for a very detailed critique of our paper, which caused us some headache but in the end led to a significantly improved presentation and a strengthening of our results. We thank Peter Harremoës for his help with the proof of Theorem 7, and Wouter Koolen for pointing out a serious error in a proof from a preliminary version of this paper. We are also grateful to Yishay Mansour, who made a single remark over lunch at COLT 2005 that sparked off this entire line of research, and to Andrew Barron, Harrison Zhou and Yuhong Yang for some very helpful conversations. An e-book version of "Alice's Adventures in Wonderland" was made available by project Gutenberg at www.gutenberg.org. This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. This publication only reflects the authors' views.

A Cumulative Risk Proofs

A.1 Oracle Approximation Lemma

The proofs of Theorems 2 and 5 both depend on the following bound on the excess cumulative risk of the switch distribution compared to any oracle.

Lemma 9 (Oracle Approximation Lemma)**.** *Let p_{sw} be the switch distribution defined with respect to a prior π (that can be written in the form (10)). Let $\mathcal{M}^*$ be a standard i.i.d. 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.*

$$m(n) = \max_{p^* \in \mathcal{M}^*} \max_{x^n \in \mathcal{X}^n} \left| \{ i : 2 \leq i \leq n, \omega(p^*, x^i) \neq \omega(p^*, x^{i-1}) \} \right| + 1. \tag{26}$$

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 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,

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

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 (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 1 to obtain the desired result. Using that π is of the form (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 1. □

A.2 Proof of Theorem 2

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

$$R_{\mathrm{m}}(p_{\omega'}, n) \preceq R_{\mathrm{m}}(p_\omega, n) \cdot \limsup_{j \to \infty} \frac{d_j}{d_{j-1}}, \tag{27}$$

where $d_j = t_{j+1} - t_j$. This construction will work for any choice of switch-points. Let $\exp_2(x) = 2^x$. Then in particular, by choosing the switch-points such that $d_j = \left\lceil \exp_2\left(j^{1/(1+\alpha)}\right) \right\rceil$, we obtain

$$\limsup_{j \to \infty} \frac{d_j}{d_{j-1}} = \limsup_{j \to \infty} \exp_2\left(\frac{j}{j^{\alpha/(1+\alpha)}} - \frac{j-1}{(j-1)^{\alpha/(1+\alpha)}} \right) \leq \limsup_{j \to \infty} \exp_2\left(\frac{1}{j^{\alpha/(1+\alpha)}} \right) = 1.$$

Let $m(n)$ denote the maximum number of different prediction strategies used by ω' before time n, as defined in (26). We must have $t_{m(n)} > n$. Hence $m(n) \leq k$ for the smallest k such that $d_k = \left\lceil \exp_2\left(k^{1/(1+\alpha)}\right) \right\rceil > n$. 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 (27) 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 (27). 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 (27). 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 $R_{\mathrm{m}}(p_\omega, n) \to \infty$, we find that

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

and (27) follows by letting ε tend to 0. □

A.3 Propositions 4 and 8

Both Proposition 4 and Proposition 8 follow from the following more general proposition.

Proposition 10. *Suppose that $\mathcal{M}^*$ is standard i.i.d. 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 with c_1, c_2, γ and h_0 as in (22), 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 8, let p be p_{sw}. To prove Proposition 4, 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 10 by letting ε tend to 0, such that c_3 tends to 1.

The proof of Proposition 10 requires the following lemma:

Lemma 11. *Let $g, h \colon \mathbb{N} \to \mathbb{R} \cup \{\infty\}$ 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

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

The lemma follows by letting ε tend to 0. □

Proof of Proposition 10. 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, we further have:

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

where the first step uses Lemma 11 and (a) follows by approximating the sum by an integral. The result follows. □

A.4 Proof of Theorem 5

The proof of Theorem 5 is based on the following lemma.

Lemma 12 (Fast Switching Lemma)**.** *Let $\mathcal{M}^*$ be standard i.i.d. and assume Condition 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 (22). 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 $f : \mathbb{N} \to [1, \infty)$, 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 (23)*.*

The proof of Lemma 12 requires the following lemma:

Lemma 13. *If $\mathcal{M}^*$ is a standard i.i.d. model class that satisfies Condition 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 $0 \leq \gamma < 1$ be constants and h_0 a nondecreasing, strictly positive function that satisfy Condition 2. Then by assumption there exists an n^* such that $r_{\mathrm{mm}}(i) \geq \frac{1}{2} c_1 i^{-\gamma} h_0(i)$ for all $i \geq n^*$. Hence

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

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

$$\sum_{j=1}^m d_j r_{\mathrm{mm}}(t_j) \geq \sum_{j=j^*}^m d_j r_{\mathrm{mm}}(t_j) \geq c_1 h_0(1) \sum_{j=j^*}^m d_j t_j^{-\gamma}. \tag{28}$$

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

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

Combining with (28) completes the proof. □

Proof of Lemma 12. Let $s(n)$ denote the last freezing time preceding n, i.e. $s(n) = t_k$ for k such that $t_k \leq n < t_{k+1}$. 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

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

As $f(t_j) \geq 1$, Lemma 13 implies that $\sum_{j=1}^m d_j f(t_j) r_{\mathrm{mm}}(t_j) \to \infty$. Therefore by Lemma 11

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

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 11 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 13, $\sum_{i=1}^n r_{\mathrm{mm}}(i) \to \infty$. Therefore by Lemma 11

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

Finally, by Proposition 4

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

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

Proof of Theorem 5. By Lemma 12 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

$$R_{\mathrm{m}}(p_{\omega'}, n) \preceq cf(n)R_{\mathrm{mm}}(n). \tag{29}$$

Let $m(n)$ denote the maximum number of different prediction strategies ω' uses before the n-th outcome, as in (26). Then the choice of $\mathbf{t}$ ensures that $m(n) = O(\log n)$, such that by the Oracle Approximation Lemma (Lemma 9) and Condition 1

$$R_{\mathrm{m}}(p_{\mathrm{sw}}, n) = R_{\mathrm{m}}(p_{\omega'}, n) + O\left((\log n)^2\right). \tag{30}$$

Finally, Proposition 4 and Condition 2 together imply that $R_{\mathrm{mm}}(n) \succeq nr_{\mathrm{mm}}(n) \succeq c_1 h_0(1) n^{1-\gamma}$, so that $(\log n)^2 / R_{\mathrm{mm}}(n) \to 0$. Combining this with (29) and (30), the result follows. □

B Consistency Proof

Proof. It is sufficient to show that

$$\lim_{n\to\infty} p_{\mathrm{sw}}(K_{n+1} \neq k^* \mid X^n) = 0 \qquad (p_{k^*}\text{-a.s.}), \tag{31}$$

which is equivalent to (25) 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 (25) 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 (31).

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 $|\mathrm{support}(\lambda_{n+1})| > 1$. (Note that $n' > n_{k^*}$ only in the degenerate case that $\lambda_n(k^*) = 1$ for all $n \leq n'$.) The assumption that $\sum_{\mathbf{s}\in B^{k^*}_{n_{k^*}}} \pi(\mathbf{s}) q_{\mathbf{s}}(X^{n_{k^*}}) > 0$ $(p_{k^*}\text{-a.s.})$ implies that

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

where $X^b_a = X_a, \ldots, X^b$. Hence the posterior distribution $\pi(\mathbf{s} \mid X^{n'}) = \frac{\pi(\mathbf{s}) q_{\mathbf{s}}(X^{n'})}{p_{\mathrm{sw}}(X^{n'})}$ is defined $(p_{k^*}\text{-a.s.})$, and by substituting definitions we find that (31) is equivalent to

$$\lim_{n\to\infty} \frac{p_{\mathrm{sw}}(X^{n'}) \sum_{\mathbf{s}\in U_n} \pi(\mathbf{s} \mid X^{n'}) q_{\mathbf{s}}(X^n_{n'+1} \mid X^{n'})}{p_{\mathrm{sw}}(X^n)} = 0 \qquad (p_{k^*}\text{-a.s.}). \tag{32}$$

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

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

which gives a bound on the numerator of (32). We may also bound the denominator by

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

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

$$\lim_{n \to \infty} \frac{r(X_{n'+1}^{n} \mid X^{n'})}{p_{k^*}(X_{n'+1}^{n} \mid X^{n'})} = 0 \qquad (p_{k^*}\text{-a.s.}), \tag{33}$$

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.

We will show that, with p_{k^*}-probability 1, the distributions $r(X_{n'+1}^{\infty} \mid X^{n'})$ and $p_{k^*}(X_{n'+1}^{\infty} \mid X^{n'})$ are mutually singular, and hence that the density ratio $r(X_{n'+1}^{\infty} \mid X^{n'}) / p_{k^*}(X_{n'+1}^{\infty} \mid X^{n'})$ is 0. As $r(X_{n'+1}^{n} \mid X^{n'}) / p_{k^*}(X_{n'+1}^{n} \mid X^{n'})$ tends to $r(X_{n'+1}^{\infty} \mid X^{n'}) / p_{k^*}(X_{n'+1}^{\infty} \mid X^{n'})$ with p_{k^*}-probability 1 (e.g. by Lévy's theorem [Shiryaev, 1996]), this implies (33).

It remains to establish mutual singularity of $r(X_{n'+1}^{\infty} \mid X^{n'})$ and $p_{k^*}(X_{n'+1}^{\infty} \mid X^{n'})$ with probability 1. To this end, we first observe that if $\mathbf{s} = \big((t_1, k_1), \ldots, (t_m, k_m)\big) \in A$ then mutual singularity of $p_{k_m}(X_{n'+1}^{\infty} \mid X^{n'})$ and $p_{k^*}(X_{n'+1}^{\infty} \mid X^{n'})$ implies mutual singularity of $q_{\mathbf{s}}(X_{n'+1}^{\infty} \mid X^{n'})$ and $p_{k^*}(X_{n'+1}^{\infty} \mid X^{n'})$. Hence $r(X_{n'+1}^{\infty} \mid X^{n'})$ is a countable mixture of distributions that are mutually singular with $p_{k^*}(X_{n'+1}^{\infty} \mid X^{n'})$ and is therefore itself mutually singular with $p_{k^*}(X_{n'+1}^{\infty} \mid X^{n'})$, as required. □

References

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 statistical model identification. *IEEE Transactions on Automatic Control*, 19 (6):716–723, 1974.

A. Barron. Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In A. D. J.M. Bernardo, J.O. Berger and A. Smith, editors, *Bayesian Statistics*, volume 6, pages 27–52. Oxford University Press, Oxford, 1998a.

A. Barron and T. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4):1034–1054, 1991.

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, 1998.

A. R. Barron. Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In *Bayesian Statistics 6*, pages 27–52. Oxford University Press, 1998b.

J. Bernardo and A. Smith. *Bayesian Theory*. Wiley, Chichester, 1994.

G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, 1973.

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.

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, 1990.

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 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.

P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *The Annals of Statistics*, 14(1): 1–26, 1986.

D. Donoho and I. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

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.

S. Ghosal, J. Lember, and A. van der Vaart. Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics*, 2:63–89, 2008.

P. D. Grünwald. *The Minimum Description Length Principle*. The MIT Press, 2007.

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.

D. Haussler and M. Opper. Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, 25(6):2451–2492, 1997.

M. Herbster and M. K. Warmuth. Tracking the best expert. *Machine Learning*, 32:151–178, 1998.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430): 773–795, 1995.

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. Published on the CoRR arXiv abs/0802.2015, 2008a.

W. Koolen and S. de Rooij. Combining expert advice efficiently. In *Proceedings of the 21st Annual Conference on Computational Learning Theory (COLT 2008)*, 2008b.

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.

J. Poland and M. Hutter. Asymptotics of discrete MDL for online prediction. *IEEE Transactions on Information Theory*, 51(11):3780–3795, 2005.

J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, IT-30(4):629–636, 1984.

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

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

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.

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. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

T. van Erven. *When Data Compression and Statistics Disagree: Two Frequentist Challenges for the Minimum Description Length Principle*. PhD thesis, Leiden University, 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 2008.

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.

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.

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.