Catching Up Faster in Bayesian Model Selection and Model Averaging

Tim van Erven **Peter Grünwald** **Steven de Rooij**
Centrum voor Wiskunde en Informatica (CWI)
Kruislaan 413, P.O. Box 94079
1090 GB Amsterdam, The Netherlands
{Tim.van.Erven,Peter.Grunwald,Steven.de.Rooij}@cwi.nl

Abstract

Bayesian model averaging, model selection and their approximations such as BIC are generally statistically consistent, but sometimes achieve slower rates of convergence than other methods such as AIC and leave-one-out cross-validation. On the other hand, these other methods can be inconsistent. We identify the *catch-up phenomenon* as a novel explanation for the slow convergence of Bayesian methods. Based on this analysis we define the switch-distribution, a modification of the Bayesian model averaging distribution. We prove that in many situations model selection and prediction based on the switch-distribution is both consistent and achieves optimal convergence rates, thereby resolving the AIC-BIC dilemma. The method is practical; we give an efficient algorithm.

1 Introduction

We consider inference based on a countable set of models (sets of probability distributions), focusing on two tasks: model selection and model averaging. In model selection tasks, the goal is to select the model that best explains the given data. In model averaging, the goal is to find the weighted combination of models that leads to the best prediction of future data from the same source.

An attractive property of some criteria for model selection is that they are consistent under weak conditions, i.e. if the true distribution P^* is in one of the models, then the P^*-probability that this model is selected goes to one as the sample size increases. BIC [14], Bayes factor model selection [8], Minimum Description Length (MDL) model selection [3] and prequential model validation [5] are examples of widely used model selection criteria that are usually consistent. However, other model selection criteria such as AIC [1] and leave-one-out cross-validation (LOO) [16], while often inconsistent, do typically yield better predictions. This is especially the case in nonparametric settings, where P^* can be arbitrarily well-approximated by a sequence of distributions in the (parametric) models under consideration, but is not itself contained in any of these. In many such cases, the predictive distribution converges to the true distribution at the optimal rate for AIC and LOO [15, 9], whereas in general BIC, the Bayes factor method and prequential validation only achieve the optimal rate to within an $O(\log n)$ factor [13, 20, 6]. In this paper we reconcile these seemingly conflicting approaches [19] by improving the rate of convergence achieved in Bayesian model selection without losing its convergence properties. First we provide an example to show why Bayes sometimes converges too slowly.

Given priors on models $\mathcal{M}_1$, $\mathcal{M}_2$, $\ldots$ and parameters therein, Bayesian inference associates each model $\mathcal{M}_k$ with the marginal distribution p_k, given in (1), obtained by averaging over the parameters according to the prior. In model selection the preferred model is the one with maximum a posteriori probability. By Bayes' rule this is $\arg\max_k p_k(x^n)w(k)$, where $w(k)$ denotes the prior probability of $\mathcal{M}_k$. We can further average over model indices, a process called Bayesian Model Averaging (BMA). The resulting distribution $p_{\text{bma}}(x^n) = \sum_k p_k(x^n)w(k)$ can be used for prediction. In a sequential setting, the probability of a data sequence $x^n := x_1, \ldots, x_n$ under a distribution p typically decreases exponentially fast in n. It is therefore common to consider $-\log p(x^n)$, which we call the *codelength* of x^n achieved by p. We take all logarithms to base 2, allowing us to measure codelength in *bits*. The name codelength refers to the correspondence between codelength functions and probability distributions based on the Kraft inequality, but one may also think of the codelength as the accumulated log loss that is incurred if we sequentially predict the x_i by conditioning on the past, i.e. using $p(\cdot|x^{i-1})$ [3, 6, 5, 11]. For BMA, we have $-\log p_{\text{bma}}(x^n) = \sum_{i=1}^{n} -\log p_{\text{bma}}(x_i|x^{i-1})$. Here the ith term represents the loss incurred when predicting x_i given x^{i-1} using $p_{\text{bma}}(\cdot|x^{i-1})$, which turns out to be equal to the posterior average: $p_{\text{bma}}(x_i|x^{i-1}) = \sum_k p_k(x_i|x^{i-1})w(k|x^{i-1})$.

Prediction using p_{bma} has the advantage that the codelength it achieves on x^n is close to the codelength of $p_{\hat{k}}$, where $\hat{k}$ is the index of best of the marginals $p_1, p_2, \ldots$ Namely, given a prior w on model indices, the difference between $-\log p_{\text{bma}}(x^n) = -\log(\sum_k p_k(x^n)w(k))$ and $-\log p_{\hat{k}}(x^n)$ must be in the range $[0, -\log w(\hat{k})]$, whatever data x^n are observed. Thus, using BMA for prediction is sensible if we are satisfied with doing essentially as well as the best model under consideration. However, it is often possible to combine $p_1, p_2, \ldots$ into a distribution that achieves smaller codelength than $p_{\hat{k}}$! This is possible if the index $\hat{k}$ of the best distribution *changes with the sample size in a predictable way*. This is common in model selection, for example with nested models, say $\mathcal{M}_1 \subset \mathcal{M}_2$. In this case p_1 typically predicts better at small sample sizes (roughly, because $\mathcal{M}_2$ has more parameters that need to be learned than $\mathcal{M}_1$), while p_2 predicts better eventually. Figure 1 illustrates this phenomenon. It shows the accumulated codelength difference $-\log p_2(x^n) - (-\log p_1(x^n))$ on "The Picture of Dorian Gray" by Oscar Wilde, where p_1 and p_2 are the Bayesian marginal distributions for the first-order and second-order Markov chains, respectively, and each character in the book is an outcome. Note that the example models $\mathcal{M}_1$ and $\mathcal{M}_2$ are very crude; for this particular application much better models are available. In more complicated, more realistic model selection scenarios, the models may still be wrong, but it may not be known how to improve them. Thus $\mathcal{M}_1$ and $\mathcal{M}_2$ serve as a simple illustration only. We used uniform priors on the model parameters, but for other common priors similar behaviour can be expected. Clearly p_1 is better for about the first $100\,000$ outcomes, gaining a head start of approximately $40\,000$ bits. Ideally we should predict the initial $100\,000$ outcomes using p_1 and the rest using p_2. However, p_{bma} only starts to behave like p_2 when it *catches up* with p_1 at a sample size of about $310\,000$, when the codelength of p_2 drops below that of p_1. Thus, in the shaded area p_{bma} behaves like p_1 while p_2 is making better predictions of those outcomes: since at $n = 100\,000$, p_2 is $40\,000$ bits behind, and at $n = 310\,000$, it has caught up, in between it must have outperformed p_1 by $40\,000$ bits!

The general pattern that first one model is better and then another occurs widely, both on real-world data and in theoretical settings. We argue that failure to take this effect into account leads to the suboptimal rate of convergence achieved by Bayes factor model selection and related methods. We have developed an alternative method to combine distributions p_1 and p_2 into a single distribution p_{sw}, which we call the *switch-distribution*, defined in Section 2. Figure 1 shows that p_{sw} behaves like p_1 initially, but in contrast to p_{bma} it starts to mimic p_2 *almost immediately* after p_2 starts making better predictions; it essentially does this *no matter what sequence x^n is actually observed*. p_{sw} differs from p_{bma} in that it is based on a prior distribution on *sequences of models* rather than simply a prior distribution on models. This allows us to avoid the implicit assumption that there is one model which is best at all sample sizes. After conditioning on past observations, the posterior we obtain gives a better indication of which model performs best *at the current sample size*, thereby achieving a faster rate of convergence. Indeed, the switch-distribution is related to earlier algorithms for *tracking the best expert* developed in the universal prediction literature [7, 18, 17, 10]; however, the applications we have in mind and the theorems we prove are completely different. In Sections 3 and 4 we show that model selection based on the switch-distribution is consistent (Theorem 1), but unlike standard

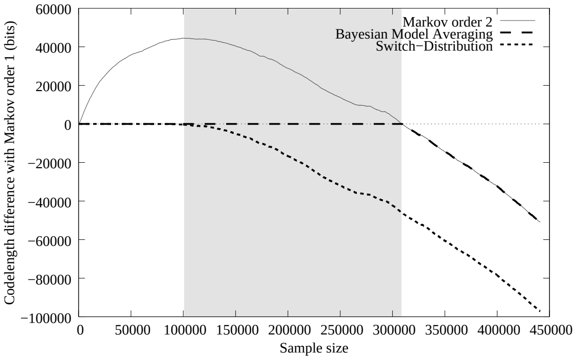

Figure 1: The Catch-up Phenomenon

Bayes factor model selection achieves optimal rates of convergence (Theorem 2). Proofs of the theorems are in Appendix A. In Section 5 we give a practical algorithm that computes the switch-distribution for K (rather than 2) predictors in $\Theta(n \cdot K)$ time. In the full paper, we will give further details of the proof of Theorem 1 and a more detailed discussion of Theorem 2 and the implications of both theorems.

2 The Switch-Distribution for Model Selection and Prediction

Preliminaries Suppose $X^\infty = (X_1, X_2, \ldots)$ is a sequence of random variables that take values in sample space $\mathcal{X} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{Z}^+ = \{1, 2, \ldots\}$. For $n \in \mathbb{N} = \{0, 1, 2, \ldots\}$, let $x^n = (x_1, \ldots, x_n)$ denote the first n outcomes of X^∞, such that x^n takes values in the product space $\mathcal{X}^n = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. (We let x^0 denote the empty sequence.) Let $\mathcal{X}^* = \bigcup_{n=0}^\infty \mathcal{X}^n$. For $m > n$, we write X_{n+1}^m for $(X_{n+1}, \ldots, X_m)$, where $m = \infty$ is allowed and we omit the subscript when $n = 0$.

Any distribution $P(X^\infty)$ may be defined by a sequential *prediction strategy* p that predicts the next outcome at any time $n \in \mathbb{N}$. To be precise: Given the previous outcomes x^n at time n, this prediction strategy should issue a conditional density $p(X_{n+1}|x^n)$ with corresponding distribution $P(X_{n+1}|x^n)$ for the next outcome X_{n+1}. Such sequential prediction strategies are sometimes called *prequential forecasting systems* [5]. An instance is given in Example 1 below. We assume that the density $p(X_{n+1}|x^n)$ 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 $p(X_{n+1}|x^n)$ is a probability mass function. It is natural to define the joint density $p(x^m|x^n) = p(x_{n+1}|x^n) \cdots p(x_m|x^{m-1})$ and let $P(X_{n+1}^\infty|x^n)$ be the unique distribution such that, for all $m > n$, $p(X_{n+1}^m|x^n)$ is the density of its marginal distribution for X_{n+1}^m. To ensure that $P(X_{n+1}^\infty|x^n)$ is well-defined even if $\mathcal{X}$ is continuous, we impose the natural requirement that for any $k \in \mathbb{Z}^+$ and any fixed event $A_{k+1} \subseteq \mathcal{X}_{k+1}$ the probability $P(A_{k+1}|x^k)$ is a measurable function of x^k, which holds automatically if $\mathcal{X}$ is countable.

Model Selection and Prediction The goal in *model selection* is to choose an explanation for observed data x^n from a potentially infinite list of candidate models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ We consider *parametric models*, which are sets $\{p_\theta : \theta \in \Theta\}$ of prediction strategies p_θ that are indexed by elements of $\Theta \subseteq \mathbb{R}^d$, for some smallest possible $d \in \mathbb{N}$, the number of degrees of freedom. Examples of model selection are regression based on a set of basis functions such as polynomials (d is the number of coefficients of the polynomial), the variable selection problem in regression [15, 9, 20] (d is the number of variables), and histogram density estimation [13] (d is the number of bins). A *model selection criterion* is a function $\delta : \mathcal{X}^* \to \mathbb{Z}^+$ that, given any data sequence $x^n \in \mathcal{X}^*$, selects the model $\mathcal{M}_k$ with index $k = \delta(x^n)$.

We associate each model $\mathcal{M}_k$ with a single prediction strategy $\bar{p}_k$. The bar emphasizes that $\bar{p}_k$ is a meta-strategy based on the prediction strategies in $\mathcal{M}_k$. In many approaches to model selection, for example AIC and LOO, $\bar{p}_k$ is defined using some estimator $\hat{\theta}_k$ for each model $\mathcal{M}_k$, which maps a sequence x^n of previous observations to an estimated parameter value that represents a "best guess" of the true/best distribution in the model. Prediction is then based on this estimator: $\bar{p}_k(X_{n+1} \mid x^n) = p_{\hat{\theta}_k(x^n)}(X_{n+1} \mid x^n)$, which also defines a joint density $\bar{p}_k(x^n) = \bar{p}_k(x_1) \cdots \bar{p}_k(x_n|x^{n-1})$. The Bayesian approach to model selection or model averaging goes the other way around. We start out with a prior w on Θ_k, and define the Bayesian marginal density

$$\bar{p}_k(x^n) = \int_{\theta \in \Theta_k} p_\theta(x^n) w(\theta) \, d\theta. \tag{1}$$

When $\bar{p}_k(x^n)$ is non-zero this joint density induces a unique conditional density $\bar{p}_k(X_{n+1} \mid x^n) = \bar{p}_k(X_{n+1}, x^n)/\bar{p}_k(x^n)$, which is equal to the mixture of $p_\theta \in \mathcal{M}_k$ according to the posterior, $w(\theta|x^n) = p_\theta(x^n)w(\theta)/\int p_\theta(x^n)w(\theta) \, d\theta$, based on x^n. Thus the Bayesian approach also defines a prediction strategy $\bar{p}_k(X_{n+1}|x^n)$, whose corresponding distribution may be thought of as an estimator. From now on we sometimes call the distributions induced by $\bar{p}_1, \bar{p}_2, \ldots$ "estimators", even if they are Bayesian. This unified view is known as *prequential* or *predictive MDL* [11, 5].

Example 1. Suppose $\mathcal{X} = \{0, 1\}$. Then a prediction strategy $\bar{p}$ may be based on the Bernoulli model $\mathcal{M} = \{p_\theta \mid \theta \in [0, 1]\}$ that regards X^∞ as a sequence of independent, identically distributed Bernoulli random variables with $P_\theta(X_{n+1} = 1) = \theta$. We may predict X_{n+1} using the maximum likelihood (ML) estimator based on the past, i.e. using $\hat{\theta}(x^n) = n^{-1} \sum_{i=1}^n x_i$. The prediction for x_1 is then undefined. If we use a smoothed ML estimator such as the Laplace estimator, $\hat{\theta}'(x^n) =$

$(n+2)^{-1}(\sum_{i=1}^{n} x_i + 1)$, then all predictions are well-defined. Perhaps surprisingly, the predictor $\bar{p}'$ defined by $\bar{p}'(X_{n+1} \mid x^n) = p_{\hat{\theta}'(x^n)}(X_{n+1})$ equals the Bayesian predictive distribution based on a uniform prior. Thus in this case a Bayesian predictor and an estimation-based predictor coincide!

The Switch-Distribution Suppose $p_1, p_2, \ldots$ is a list of prediction strategies for X^∞. (Although here the list is infinitely long, the developments below can with little modification be adjusted to the case where the list is finite.) We first define a family $\mathcal{Q} = \{q_\mathbf{s} : \mathbf{s} \in \mathbb{S}\}$ of combinator prediction strategies that switch between the original prediction strategies. Here the parameter space $\mathbb{S}$ is defined as

$$\mathbb{S} = \{(t_1, k_1), \ldots, (t_m, k_m) \in (\mathbb{N} \times \mathbb{Z}^+)^m \mid m \in \mathbb{Z}^+, 0 = t_1 < \ldots < t_m\}. \tag{2}$$

The parameter $\mathbf{s} \in \mathbb{S}$ specifies the identities of m constituent prediction strategies and the sample sizes, called *switch-points*, at which to switch between them. For $\mathbf{s} = ((t'_1, k'_1), \ldots, (t'_{m'}, k'_{m'}))$, we define $t_i(\mathbf{s}) = t'_i$, $k_i(\mathbf{s}) = k'_i$ and $m(\mathbf{s}) = m'$. We omit the argument when the parameter $\mathbf{s}$ is clear from context, e.g. we write t_3 for $t_3(\mathbf{s})$. For each $\mathbf{s} \in \mathbb{S}$ the corresponding $q_\mathbf{s} \in \mathcal{Q}$ is defined as:

$$q_\mathbf{s}(X_{n+1}|x^n) = \begin{cases} p_{k_1}(X_{n+1}|x^n) & \text{if } n < t_2, \\ p_{k_2}(X_{n+1}|x^n) & \text{if } t_2 \le n < t_3, \\ \vdots & \vdots \\ p_{k_{m-1}}(X_{n+1}|x^n) & \text{if } t_{m-1} \le n < t_m, \\ p_{k_m}(X_{n+1}|x^n) & \text{if } t_m \le n. \end{cases} \tag{3}$$

Switching to the same predictor multiple times is allowed. The extra switch-point t_1 is included to simplify notation; we always take $t_1 = 0$. Now the switch-distribution is defined as a Bayesian mixture of the elements of $\mathcal{Q}$ according to a prior π on $\mathbb{S}$:

Definition 1 (Switch-Distribution)**.** *Let π be a probability mass function on $\mathbb{S}$. Then the switch-distribution P_{sw} with prior π is the distribution for X^∞ such that, for any $n \in \mathbb{Z}^+$, the density of its marginal distribution for X^n is given by*

$$p_{\text{sw}}(x^n) = \sum_{\mathbf{s} \in \mathbb{S}} q_\mathbf{s}(x^n) \cdot \pi(\mathbf{s}). \tag{4}$$

Although the switch-distribution provides a general way to combine prediction strategies, in this paper it will only be applied to combine prediction strategies $\bar{p}_1, \bar{p}_2, \ldots$ that correspond to models. In this case we may define a corresponding model selection criterion δ_{sw}. To this end, let $K_{n+1} : \mathbb{S} \to \mathbb{Z}^+$ be a random variable that denotes the strategy/model that is used to predict X_{n+1} given past observations x^n. Formally, $K_{n+1}(\mathbf{s}) = k_i(\mathbf{s})$ iff $t_i(\mathbf{s}) \le n$ and $i = m(\mathbf{s}) \vee n < t_{i+1}(\mathbf{s})$. Algorithm 1, given in Section 5, efficiently computes the posterior distribution on K_{n+1} given x^n:

$$\pi(K_{n+1} = k \mid x^n) = \frac{\sum_{\{\mathbf{s}: K_{n+1}(\mathbf{s}) = k\}} \pi(\mathbf{s}) q_\mathbf{s}(x^n)}{p_{\text{sw}}(x^n)}, \tag{5}$$

which is defined whenever $p_{\text{sw}}(x^n)$ is non-zero. We turn this into a model selection criterion $\delta_{\text{sw}}(x^n) = \arg\max_k \pi(K_{n+1} = k|x^n)$ that selects the model with maximum posterior probability.

3 Consistency

If one of the models, say with index k^*, is actually true, then it is natural to ask whether δ_{sw} is *consistent*, in the sense that it asymptotically selects k^* with probability 1. Theorem 1 below states that this is the case under certain conditions which are only slightly stronger than those required for the consistency of standard Bayes factor model selection.

Bayes factor model selection is consistent if for all $k, k' \neq k$, $\bar{P}_k(X^\infty)$ and $\bar{P}_{k'}(X^\infty)$ are mutually singular, that is, if there exists a measurable set $A \subseteq \mathcal{X}^\infty$ such that $\bar{P}_k(A) = 1$ and $\bar{P}_{k'}(A) = 0$ [3]. For example, this can usually be shown to hold if the models are nested and for each k, Θ_k is a subset of Θ_{k+1} of w_{k+1}-measure 0 [6]. For consistency of δ_{sw}, we need to strengthen this to the requirement that, for all $k' \neq k$ and all $x^n \in \mathcal{X}^*$, the distributions $\bar{P}_k(X_{n+1}^\infty \mid x^n)$ and $\bar{P}_{k'}(X_{n+1}^\infty \mid x^n)$ are mutually singular. For example, if $X_1, X_2, \ldots$ are i.i.d. according to each P_θ in all models, but also if $\mathcal{X}$ is countable and $\bar{p}_k(x_{n+1} \mid x_n) > 0$ for all k, all $x^{n+1} \in \mathcal{X}^{n+1}$, then this conditional mutual singularity is automatically implied by ordinary mutual singularity of $\bar{P}_k(X^\infty)$ and $\bar{P}_{k'}(X^\infty)$.

Let $E_{\mathbf{s}} = \{\mathbf{s}' \in \mathbb{S} \mid m(\mathbf{s}') > m(\mathbf{s}), (t_i(\mathbf{s}'), k_i(\mathbf{s}')) = (t_i(\mathbf{s}), k_i(\mathbf{s})) \text{ for } i = 1, \ldots, m(\mathbf{s})\}$ denote the set of all possible extensions of $\mathbf{s}$ to more switch-points. Let $\bar{p}_1, \bar{p}_2, \ldots$ be Bayesian prediction strategies with respective parameter spaces $\Theta_1, \Theta_2, \ldots$ and priors $w_1, w_2, \ldots$, and let π be the prior of the corresponding switch-distribution.

Theorem 1 (Consistency of the Switch-Distribution)**.** *Suppose π is positive everywhere on $\{\mathbf{s} \in \mathbb{S} \mid m(\mathbf{s}) = 1\}$ and is such that there exists a positive constant c such that, for every $\mathbf{s} \in \mathbb{S}$, $c \cdot \pi(\mathbf{s}) \geq \pi(E_{\mathbf{s}})$. Suppose further that $\bar{P}_k(X_{n+1}^{\infty} \mid x^n)$ and $\bar{P}_{k'}(X_{n+1}^{\infty} \mid x^n)$ are mutually singular for all $k, k' \in \mathbb{Z}^+$, $k \neq k'$, $x^n \in \mathcal{X}^*$. Then, for all $k^* \in \mathbb{Z}^+$, 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*

$$\pi(K_{n+1} = k^* \mid X^n) \overset{n\to\infty}{\longrightarrow} 1 \qquad \textit{with } P_{\theta^*}\textit{-probability } 1. \tag{6}$$

The requirement that $c \cdot \pi(\mathbf{s}) \geq \pi(E_{\mathbf{s}})$ is automatically satisfied if π is of the form:

$$\pi(\mathbf{s}) = \pi_{\text{M}}(m)\pi_{\text{K}}(k_1) \prod_{i=2}^{m} \pi_{\text{T}}(t_i | t_i > t_{i-1})\pi_{\text{K}}(k_i), \tag{7}$$

where π_{M}, π_{K} and π_{T} are priors on $\mathbb{Z}^+$ with full support, and π_{M} is geometric: $\pi_{\text{M}}(m) = \theta^{m-1}(1-\theta)$ for some $0 \leq \theta < 1$. In this case $c = \theta/(1-\theta)$.

4 Optimal Risk Convergence Rates

Suppose $X_1, X_2, \ldots$ are distributed according to P^*. We define the *risk* at sample size $n \geq 1$ of the estimator $\bar{P}$ relative to P^* as

$$R_n(P^*, \bar{P}) = E_{X^{n-1} \sim P^*}[D(P^*(X_n = \cdot \mid X^{n-1}) \| \bar{P}(X_n = \cdot \mid X^{n-1}))],$$

where $D(\cdot \| \cdot)$ is the Kullback-Leibler (KL) divergence [4]. This is the standard definition of risk relative to KL divergence. The risk is always well-defined, and equal to 0 if $\bar{P}(X_{n+1} \mid X^n)$ is equal to $P^*(X_{n+1} \mid X^n)$. The following identity connects information-theoretic redundancy and accumulated statistical risk (see [4] or [6, Chapter 15]): If P^* admits a density p^*, then for all prediction strategies $\bar{p}$,

$$E_{X^n \sim P^*}[-\log \bar{p}(X^n) + \log p^*(X^n)] = \sum_{i=1}^{n} R_i(P^*, \bar{P}). \tag{8}$$

For a union of parametric models $\mathcal{M} = \bigcup_{k \geq 1} \mathcal{M}_k$, we define the *information closure* $\langle \mathcal{M} \rangle = \{P^* \mid \inf_{P \in \mathcal{M}} D(P^* \| P) = 0\}$, i.e. the set of distributions for X^∞ that can be arbitrarily well approximated by elements of $\mathcal{M}$. Theorem 2 below shows that, for a very large class of $P^* \in \langle \mathcal{M} \rangle$, the switch-distribution defined relative to estimators $\bar{P}_1, \bar{P}_2, \ldots$ achieves the same risk as any other model selection criterion defined with respect to the same estimators, up to lower order terms; in other words, model averaging based on the switch-distribution achieves at least the same rate of convergence as model selection based on any model selection criterion whatsoever (the issue of averaging vs selection will be discussed at length in the full paper). The theorem requires that the prior π in (4) is of the form (7), and satisfies

$$-\log \pi_{\text{M}}(m) = O(m) \;\; ; \;\; -\log \pi_{\text{K}}(k) = O(\log k) \;\; ; \;\; -\log \pi_{\text{T}}(t) = O(\log t). \tag{9}$$

Thus, π_{M}, the prior on the total number of switch points, is allowed to decrease either polynomially or exponentially (as required for Theorem 1); π_{T} and π_{K} must decrease polynomially. For example, we could set $\pi_{\text{T}}(t) = \pi_{\text{K}}(t) = 1/(t(t+1))$, or we could take the universal prior on the integers [12].

Let $\mathcal{M}^* \subset \langle \mathcal{M} \rangle$ be some subset of interest of the information closure of model $\mathcal{M}$. $\mathcal{M}^*$ may consist of just a single, arbitrary distribution P^* in $\langle \mathcal{M} \rangle \setminus \mathcal{M}$ – in that case Theorem 2 shows that the switch-distribution converges as fast as any other model selection criterion on any distribution in $\langle \mathcal{M} \rangle$ that cannot be expressed parametrically relative to $\mathcal{M}$ – or it may be a large, nonparametric family. In that case, Theorem 2 shows that the switch-distribution achieves the minimax convergence rate. For example, if the models $\mathcal{M}_k$ are k-bin histograms [13], then $\langle \mathcal{M} \rangle$ contains every distribution on $[0,1]$ with bounded continuous densities, and we may, for example, take $\mathcal{M}^*$ to be the set of all distributions on $[0,1]$ which have a differentiable density p^* such that $p^*(x)$ and $(\mathrm{d}/\mathrm{d}x)p^*(x)$ are bounded from below and above by some positive constants.

We restrict ourselves to model selection criteria which, at sample size n, never select a model $\mathcal{M}_k$ with $k > n^\tau$ for some arbitrarily large but fixed $\tau > 0$; note that this condition will be met for most practical model selection criteria. Let $h : \mathbb{Z}^+ \to \mathbb{R}^+$ denote the minimax optimal achievable risk as a function of the sample size, i.e.

$$h(n) = \inf_{\delta:\mathcal{X}^n \to \{1,2,\ldots,\lceil n^\tau \rceil\}} \sup_{P^* \in \mathcal{M}^*} \sup_{n' \geq n} R_{n'}(P^*, \bar{P}_\delta), \tag{10}$$

where the infimum is over all model selection criteria restricted to sample size n, and $\lceil \cdot \rceil$ denotes rounding up to the nearest integer. $\bar{p}_\delta$ is the prediction strategy satisfying, for all $n' \geq n$, all $x^{n'} \in \mathcal{X}^{n'}$, $\bar{p}_\delta(X_{n'+1} \mid x^{n'}) := \bar{p}_{\delta(x^n)}(X_{n'+1} \mid x^{n'})$, i.e. at sample size n it predicts x_{n+1} using $\bar{p}_k$ for the $k = \delta(X^n)$ chosen by δ, and it keeps predicting future $x_{n'+1}$ by this k. We call $h(n)$ the minimax optimal rate of convergence for model selection relative to data from $\mathcal{M}^*$, model list $\mathcal{M}_1, \mathcal{M}_2, \ldots$, and estimators $\bar{P}_1, \bar{P}_2, \ldots$ The definition is slightly nonstandard, in that we require a second supremum over $n' \geq n$. This is needed because, as will be discussed in the full paper, it can sometimes happen that, for some P^*, some k, some $n' > n$, $R_{n'}(P^*, \bar{P}_k) > R_n(P^*, \bar{P}_k)$ (see also [4, Section 7.1]). In cases where this cannot happen, such as regression with standard ML estimators, and in cases where, uniformly for all k, $\sup_{n' \geq n} R_{n'}(P^*, \bar{P}_k) - R_n(P^*, \bar{P}_k) = o(\sum_{i=1}^n h(i))$ (in the full paper we show that this holds for, for example, histogram density estimation), our Theorem 2 also implies minimax convergence in terms of the standard definition, without the $\sup_{n' \geq n}$. We expect that the $\sup_{n' \geq n}$ can be safely ignored for most "reasonable" models and estimators.

Theorem 2. *Define P_sw for some model class $\mathcal{M} = \cup_{k \geq 1} \mathcal{M}_k$ as in* (4)*, where the prior π satisfies (9). Let $\mathcal{M}^*$ be a subset of $\langle \mathcal{M} \rangle$ with minimax rate h such that $nh(n)$ is increasing, and $nh(n)/(\log n)^2 \to \infty$. Then*

$$\limsup_{n \to \infty} \frac{\sup_{P^* \in \mathcal{M}^*} \sum_{i=1}^n R_i(P^*, P_\text{sw})}{\sum_{i=1}^n h(i)} \leq 1. \tag{11}$$

The requirement that $nh(n)/(\log n)^2 \to \infty$ will typically be satisfied whenever $\mathcal{M}^* \setminus \mathcal{M}$ is nonempty. Then $\mathcal{M}^*$ contains P^* that are "nonparametric" relative to the chosen sequence of models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ Thus, the problem should not be "too simple": we do not know whether (11) holds in the parametric setting where $P^* \in \mathcal{M}_k$ for some k on the list. Theorem 2 expresses that the *accumulated risk* of the switch-distribution, as n increases, is not significantly larger than the *accumulated risk* of any other procedure. This "convergence in sum" has been considered before by, for example, [13, 4], and is compared to ordinary convergence in the full paper, where we will also give example applications of the theorem and further discuss (10). The proof works by bounding the redundancy of the switch-distribution, which, by (8), is identical to the accumulated risk. It is not clear whether similar techniques can be used to bound the individual risk.

5 Computing the Switch-Distribution

Algorithm 1 sequentially computes the posterior probability on predictors $p_1, p_2, \ldots$. It requires that π is a prior of the form in (7), and π_M is geometric, as is also required for Theorem 1 and permitted in Theorem 2. The algorithm resembles FIXED-SHARE [7], but whereas FIXED-SHARE implicitly imposes a geometric distribution for π_T, we allow general priors by varying the shared weight with n. We do require slightly more space to cope with π_M.

Algorithm 1 SWITCH(x^N)

```
    ▷ $K$ is the number of experts; $\theta$ is as in the definition of $\pi_\mathrm{M}$.
    for $k=1,\ldots,K$ do initialise $w_k^a \leftarrow \theta \cdot \pi_\mathrm{K}(k)$; $w_k^b \leftarrow (1-\theta) \cdot \pi_\mathrm{K}(k)$ od
    Report prior $\pi(K_1) = w_{K_1}^a$ (a $K$-sized array)
    for $n=1,\ldots,N$ do
        for $k=1,\ldots,K$ do $w_k^a \leftarrow w_k^a \cdot p_k(x_n|x^{n-1})$; $w_k^b \leftarrow w_k^b \cdot p_k(x_n|x^{n-1})$ od    (loss update)
        $\mathtt{pool} \leftarrow \pi_\mathrm{T}(Z = n \mid Z \geq n) \cdot \sum_k w_k^a$                                    (share update)
        for $k=1,\ldots,K$ do
            $w_k^a \leftarrow w_k^a \cdot \pi_\mathrm{T}(Z \neq n \mid Z \geq n)$   +          $\theta \cdot \mathtt{pool} \cdot \pi_\mathrm{K}(k)$
            $w_k^b \leftarrow w_k^b$                                +   $(1-\theta) \cdot \mathtt{pool} \cdot \pi_\mathrm{K}(k)$
        od
        Report posterior $\pi(K_{n+1} \mid x^n) = (w_{K_{n+1}}^a + w_{K_{n+1}}^b)/\sum_k (w_k^a + w_k^b)$    (a $K$-sized array)
    od
```

This algorithm can be used to obtain fast convergence in the sense of Theorem 2, which can be extended to cope with a restriction to only the first K experts. Theorem 1 can be extended to show consistency in this case as well. If $\pi_{\mathrm{T}}(Z = n \mid Z \geq n)$ and $\pi_{\mathrm{K}}(k)$ can be computed in constant time, then the running time is $\Theta(N \cdot K)$, which is of the same order as that of fast model selection criteria like AIC and BIC. We will explain this algorithm in more detail in a forthcoming publication.

Acknowledgements We thank Y. Mansour, whose remark over lunch at COLT 2005 sparked off all this research. We thank P. Harremoës and W. Koolen for mathematical support. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

A Proofs

Proof of Theorem 1. Let $U_n = \{\mathbf{s} \in \mathbb{S} \mid K_{n+1}(\mathbf{s}) \neq k^*\}$ denote the set of 'bad' parameters $\mathbf{s}$ that select an incorrect model. It is sufficient to show that

$$\lim_{n\to\infty} \frac{\sum_{\mathbf{s}\in U_n} \pi(\mathbf{s}) q_{\mathbf{s}}(X^n)}{\sum_{\mathbf{s}\in\mathbb{S}} \pi(\mathbf{s}) q_{\mathbf{s}}(X^n)} = 0 \qquad \text{with } \bar{P}_{k^*}\text{-probability 1.} \tag{12}$$

To see this, suppose the theorem is false. Then there exists a $\Phi \subseteq \Theta_{k^*}$ with $w_{k^*}(\Phi) > 0$ such that (6) does not hold for any $\theta^* \in \Phi$. But then by definition of $\bar{P}_{k^*}$ we have a contradiction with (12). Now let $A = \{\mathbf{s} \in \mathbb{S} : k_m(\mathbf{s}) \neq k^*\}$ denote the set of parameters that are bad for sufficiently large n. We observe that for each $\mathbf{s}' \in U_n$ there exists at least one element $\mathbf{s} \in A$ that uses the same sequence of switch-points and predictors on the first $n+1$ outcomes (this implies that $K_i(\mathbf{s}) = K_i(\mathbf{s}')$ for $i = 1, \ldots, n+1$) and has no switch-points beyond n (i.e. $t_m(\mathbf{s}) \leq n$). Consequently, either $\mathbf{s}' = \mathbf{s}$ or $\mathbf{s}' \in E_{\mathbf{s}}$. Therefore

$$\sum_{\mathbf{s}'\in U_n} \pi(\mathbf{s}') q_{\mathbf{s}'}(x^n) \;\leq\; \sum_{\mathbf{s}\in A} \left(\pi(\mathbf{s}) + \pi(E_{\mathbf{s}})\right) q_{\mathbf{s}}(x^n) \;\leq\; (1+c) \sum_{\mathbf{s}\in A} \pi(\mathbf{s}) q_{\mathbf{s}}(x^n). \tag{13}$$

Defining the mixture $r(x^n) = \sum_{\mathbf{s}\in A} \pi(\mathbf{s}) q_{\mathbf{s}}(x^n)$, we will show that

$$\lim_{n\to\infty} \frac{r(X^n)}{\pi(\mathbf{s} = (0, k^*)) \cdot \bar{p}_{k^*}(X^n)} = 0 \qquad \text{with } \bar{P}_{k^*}\text{-probability 1.} \tag{14}$$

Using (13) and the fact that $\sum_{\mathbf{s}\in\mathbb{S}} \pi(\mathbf{s}) q_{\mathbf{s}}(x^n) \geq \pi(\mathbf{s} = (0, k^*)) \cdot \bar{p}_{k^*}(x^n)$, this implies (12). For all $\mathbf{s} \in A$ and $x^{t_m(\mathbf{s})} \in \mathcal{X}^{t_m(\mathbf{s})}$, by definition $Q_{\mathbf{s}}(X_{t_m+1}^{\infty} | x^{t_m})$ equals $\bar{P}_{k_m}(X_{t_m+1}^{\infty} | x^{t_m})$, which is mutually singular with $\bar{P}_{k^*}(X_{t_m+1}^{\infty} | x^{t_m})$ by assumption. If $\mathcal{X}$ is a separable metric space, which holds because $\mathcal{X} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{Z}^+$, it can be shown that this conditional mutual singularity implies mutual singularity of $Q_{\mathbf{s}}(X^{\infty})$ and $\bar{P}_{k^*}(X^{\infty})$. To see this for countable $\mathcal{X}$, let $B_{x^{t_m}}$ be any event such that $Q_{\mathbf{s}}(B_{x^{t_m}} | x^{t_m}) = 1$ and $\bar{P}_{k^*}(B_{x^{t_m}} | x^{t_m}) = 0$. Then, for $B = \{y^{\infty} \in \mathcal{X}^{\infty} \mid y_{t_m+1}^{\infty} \in B_{y^{t_m}}\}$, we have that $Q_{\mathbf{s}}(B) = 1$ and $\bar{P}_{k^*}(B) = 0$. In the uncountable case, however, B may not be measurable. We omit the full proof, which was shown to us by P. Harremoës. Any countable mixture of distributions that are mutually singular with P_{k^*}, in particular R, is mutually singular with P_{k^*}. This implies (14) by Lemma 3.1 of [2], which says that for any two mutually singular distributions R and P, the density ratio $r(X^n)/p(X^n)$ goes to 0 as $n \to \infty$ with P-probability 1. □

Proof of Theorem 2. We will show that for every $\alpha > 1$,

$$\sup_{P^*\in\mathcal{M}^*} \sum_{i=1}^{n} R_i(P^*, P_{\text{sw}}) \leq \alpha \sum_{i=1}^{n} h(i) + \epsilon_{\alpha,n} \sum_{i=1}^{n} h(i), \tag{15}$$

where $\epsilon_{\alpha,n} \overset{n\to\infty}{\longrightarrow} 0$, and $\epsilon_{\alpha,1}, \epsilon_{\alpha,2}, \ldots$ are fixed constants that only depend on α, but not on the chosen subset $\mathcal{M}^*$ of $\langle\mathcal{M}\rangle$. Theorem 2 is a consequence of (15), which we will proceed to prove. Let $\delta_n : \mathcal{X}^n \to \{1, \ldots, \lceil n^{\tau} \rceil\}$ be a model selection criterion, restricted to samples of size n, that is minimax optimal, i.e. it achieves the infimum in (10). If such a δ_n does not exist, we take a δ_n that is almost minimax optimal in the sense that it achieves the infimum to within $h(n)/n$. For $j \geq 1$, let $t_j = \lceil \alpha^{j-1} \rceil - 1$. Fix an arbitrary $n > 0$ and let m be the unique integer such that $t_m < n \leq t_{m+1}$. We will first show that for arbitrary x^n, p_{sw} achieves redundancy not much worse than $q_{\mathbf{s}}$ with $\mathbf{s} = (t_1, k_1), \ldots, (t_m, k_m)$, where $k_i = \delta_{t_i}(x^{t_i})$. Then we show that the redundancy of this $q_{\mathbf{s}}$ is small enough for (15) to hold. Thus, to achieve this redundancy, it is sufficient to take only a logarithmic number $m - 1$ of switch-points: $m - 1 < \log_{\alpha}(n+1)$. Formally, we have, for some $c > 0$, uniformly for all n, $x^n \in \mathcal{X}^n$,

$$-\log p_{\text{sw}}(x^n) = -\log \sum_{\mathbf{s}' \in \mathbb{S}} q_{\mathbf{s}'}(x^n)\pi(\mathbf{s}') \leq -\log q_{\mathbf{s}}(x^n) - \log \pi_{\text{M}}(m) - \sum_{j=1}^{m} \log \pi_{\text{T}}(t_j)\pi_{\text{K}}(k_j)$$

$$\leq -\log q_{\mathbf{s}}(x^n) + c\log(n+1) + cm(\tau+1)\log n = -\log q_{\mathbf{s}}(x^n) + O((\log n)^2). \quad (16)$$

Here the second inequality follows because of (9), and the final equality follows because $m \leq \log_\alpha(n+1)+1$. Now fix any $P^* \in \langle \mathcal{M} \rangle$. Since $P^* \in \langle \mathcal{M} \rangle$, it must have some density p^*. Thus, applying (8), and then (16), and then (8) again, we find that

$$\sum_{i=1}^{n} R_i(P^*, P_{\text{sw}}) = E_{X^n \sim P^*}[-\log p_{\text{sw}}(X^n) + \log p^*(X^n)]$$

$$\leq E_{X^n \sim P^*}[-\log q_{\mathbf{s}}(X^n) + \log p^*(X^n)] + O((\log n)^2)$$

$$= \sum_{i=1}^{n} R_i(P^*, Q_{\mathbf{s}}) + O((\log n)^2) = \sum_{j=1}^{m} \sum_{i=t_j+1}^{\min\{t_{j+1},n\}} R_i(P^*, \bar{P}_{k_j}) + O((\log n)^2). \quad (17)$$

For i appearing in the second sum, with $t_j < i \leq t_{j+1}$, we have $R_i(P^*, \bar{P}_{k_j}) \leq \sup_{i' \geq t_{j+1}} R_{i'}(P^*, \bar{P}_{k_j}) = \sup_{i' \geq t_{j+1}} R_{i'}(P^*, \bar{P}_{\delta_{t_j}(x^{t_j})}) \leq h(t_j+1)$, so that

$$R_i(P^*, \bar{P}_{k_j}) \leq \frac{1}{t_j+1} \cdot (t_j+1)h(t_j+1) \leq \frac{1}{t_j+1} \cdot ih(i) \leq \frac{t_{j+1}}{t_j+1} h(i) \leq \alpha h(i),$$

where the middle inequality follows because $nh(n)$ is increasing (condition (b) of the theorem). Summing over i, we get $\sum_{j=1}^{m} \sum_{i=t_j+1}^{\min\{t_{j+1},n\}} R_i(P^*, \bar{P}_{k_j}) \leq \alpha \sum_{i=1}^{n} h(i)$. Combining this with (17), it follows that $\sum_{i=1}^{n} R_i(P^*, P_{\text{sw}}) \leq \alpha \sum_{i=1}^{n} h(i) + O((\log n)^2)$. Because this holds for arbitrary $P^* \in \mathcal{M}^*$ (with the constant in the O notation not depending on P^*), (15) now follows by the requirement of Theorem 2 that $nh(n)/(\log n)^2 \to \infty$. □

References

[1] H. Akaike. A new look at statistical model identification. *IEEE T. Automat. Contr.*, 19(6):716–723, 1974.

[2] A. Barron. *Logically Smooth Density Estimation*. PhD thesis, Stanford University, Stanford, CA, 1985.

[3] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE T. Inform. Theory*, 44(6):2743–2760, 1998.

[4] 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, 1998.

[5] A. P. Dawid. Statistical theory: The prequential approach. *J. Roy. Stat. Soc. A*, 147, Part 2:278–292, 1984.

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

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

[8] R. E. Kass and A. E. Raftery. Bayes factors. *J. Am. Stat. Assoc.*, 90(430):773–795, 1995.

[9] K. Li. Asymptotic optimality of c_p , c_l, cross-validation and generalized cross-validation: Discrete index set. *Ann. Stat.*, 15:958–975, 1987.

[10] C. Monteleoni and T. Jaakkola. Online learning of non-stationary sequences. In *Advances in Neural Information Processing Systems*, volume 16, Cambridge, MA, 2004. MIT Press.

[11] J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE T. Inform. Theory*, IT-30(4): 629–636, 1984.

[12] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, 1989.

[13] J. Rissanen, T. P. Speed, and B. Yu. Density estimation by stochastic complexity. *IEEE T. Inform. Theory*, 38(2):315–323, 1992.

[14] G. Schwarz. Estimating the dimension of a model. *Ann. Stat.*, 6(2):461–464, 1978.

[15] R. Shibata. Asymptotic mean efficiency of a selection of regression variables. *Ann. I. Stat. Math.*, 35: 415–423, 1983.

[16] M. Stone. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Stat. Soc. B*, 39:44–47, 1977.

[17] 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.

[18] V. Vovk. Derandomizing stochastic prediction strategies. *Machine Learning*, 35:247–282, 1999.

[19] Y. Yang. Can the strengths of AIC and BIC be shared? *Biometrica*, 92(4):937–950, 2005.

[20] Y. Yang. Model selection for nonparametric regression. *Statistica Sinica*, 9:475–499, 1999.