Minimum Description Length Model Selection

de Rooij, S.

Publication date
2008

Chapter 5

Slow Convergence: the Catch-up Phenomenon

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.

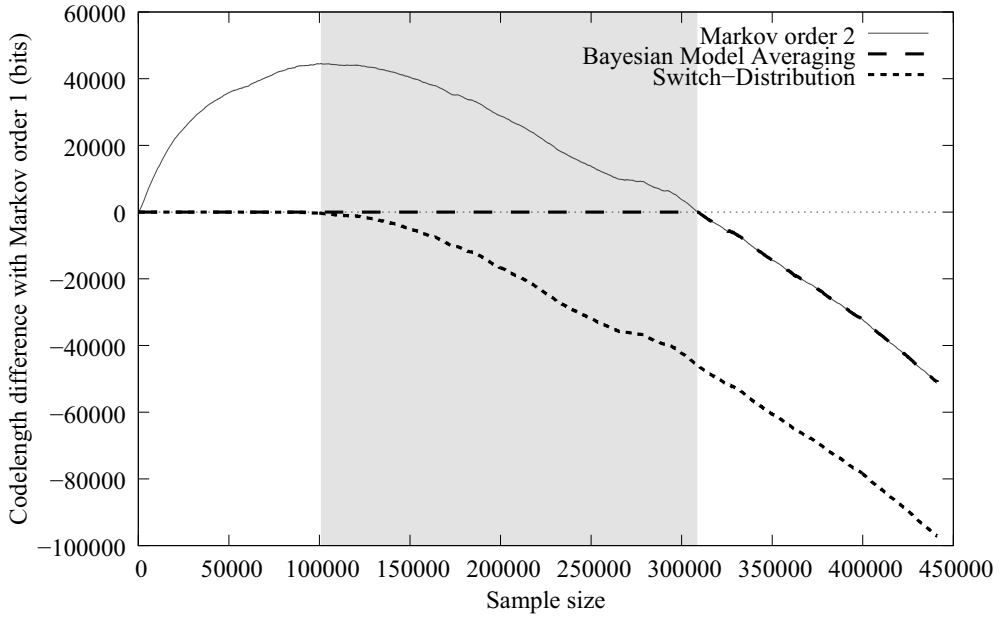
An attractive property of some criteria for model selection is that they are consistent under weak conditions, i.e. if the true distribution P^* is in one of the models, then the P^*-probability that this model is selected goes to one as the sample size increases. BIC [78], Bayes factor model selection [49], Minimum Description Length (MDL) model selection [4] and prequential model validation [27] are examples of widely used model selection criteria that are usually consistent. However, other model selection criteria such as AIC [2] and leave-one-out cross-validation (LOO) [86], while often inconsistent, do typically yield better predictions. This is especially the case in nonparametric settings of the following type: 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 [80, 56], whereas in general BIC, the Bayes factor method and prequential validation only achieve the optimal rate to within an $O(\log n)$ factor [74, 34, 101, 39]. In this paper we reconcile these seemingly conflicting approaches [103] by improving the rate of convergence achieved in Bayesian model selection without losing its consistency properties. First we provide an example to show why Bayes sometimes converges too slowly.

Given priors on models $\mathcal{M}_1$, $\mathcal{M}_2$, ... and parameters therein, Bayesian inference associates each model $\mathcal{M}_k$ with the marginal distribution p_k, given in (5.1), obtained by averaging over the parameters according to the prior. In Bayes factor 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 *code length* of x^n achieved by p. We take all logarithms to base 2, allowing us to measure code length in *bits*. The name code length refers to the correspondence between code length functions and probability distributions based on the Kraft inequality, but one may also think of the code length as the accumulated log loss that is incurred if we sequentially predict the x_i by conditioning on the past, i.e. using $p(\cdot|x^{i-1})$ [4, 39, 27, 69]. 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 code length it achieves on x^n is close to the code length of $p_{\hat{k}}$, where $\hat{k}$ is the best of the marginals $p_1, p_2, \ldots$, i.e. $\hat{k}$ achieves $\min_k -\log p_k(x^n)$. More precisely, 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 code length 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 5.1 illustrates this phenomenon. It shows the accumulated code length 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. However, 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 (see the discussion in Section 5.7.1). 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 code length 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:

Figure 5.1 The Catch-up Phenomenon



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 $p_{\rm sw}$, which we call the *switch-distribution*, defined in Section 5.1. Figure 5.1 shows that $p_{\rm sw}$ behaves like p_1 initially, but in contrast to $p_{\rm bma}$ 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*. $p_{\rm sw}$ differs from $p_{\rm bma}$ 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 very closely related to earlier algorithms for *tracking the best expert* developed in the universal prediction literature; see also Section 5.6 [46, 95, 94, 63]; however, the applications we have in mind and the theorems we prove are completely different.

The remainder of the paper is organised as follows. In Section 5.1 we introduce our basic concepts and notation, and we then define the switch-distribution. While in the example above, we switched just between two models, the general definition allows switching between elements of any finite or countably infinite

set of models. In Section 5.2 and 5.3 we show that model selection based on the switch-distribution is consistent (Theorem 5.2.1). Then in Section 5.3 we show that the switch-distribution achieves a rate of convergence that is never significantly worse than that of Bayesian model averaging, and we develop a number of tools that can be used to bound the rate of convergence "in sum" compared to other model selection criteria. Using these results we show that, in particular, the switch-distribution achieves the *worst-case optimal* rate of convergence when it is applied to histogram density estimation. In Section 5.4 we provide additional discussion, where we compare our "in sum" convergence rates to the standard definition of convergence rates, and where we motivate our conjecture that the switch-distribution in fact achieves the worst-case optimal rate of convergence in a very wide class of problems including regression using mean squared error. In Section 5.5 we give a practical algorithm that computes the switch-distribution. Theorem 5.5.1 shows that the run-time for k predictors is $\Theta(n \cdot k)$ time. In Sections 5.6 and Section 5.7 we put our work in a broader context and explain how our results fit into the existing literature. Specifically, Section 5.7.1 describes a strange implication of the catch-up phenomenon for Bayes factor model selection. The proofs of all theorems are in Section 5.9.

5.1 The Switch-Distribution for Model Selection and Prediction

5.1.1 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. We sometimes omit the subscript when $n = 0$ and write X^m rather than X_0^m.

Any distribution $P(X^\infty)$ may be defined in terms of 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* [27]. An instance is given in Example 14 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 on $\mathcal{X}^\infty$ 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 will only allow prediction strategies satisfying the natural requirement that for any $k \in \mathbb{Z}^+$ and any fixed measurable event $A_{k+1} \subseteq \mathcal{X}_{k+1}$ the probability $P(A_{k+1}|x^k)$ is a measurable function of x^k. This requirement holds automatically if $\mathcal{X}$ is countable.

5.1.2 Model Selection and Prediction

In *model selection* the goal is to choose an explanation for observed data x^n from a potentially infinite list of candidate models $\mathcal{M}_1$, $\mathcal{M}_2$, ... We consider *parametric models*, which we define as 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. A model is more commonly viewed as a set of distributions, but since distributions can be viewed as prediction strategies as explained above, we may think of a model as a set of prediction strategies as well. 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 [80, 56, 101] (d is the number of variables), and histogram density estimation [74] (d is the number of bins minus 1). 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 $k = \delta(x^n)$.

With each model $\mathcal{M}_k$ we associate a single prediction strategy $\bar{p}_k$. The bar emphasises 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 $\hat{\theta}_k$, 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. It starts out with a prior w on Θ_k, and then defines the Bayesian marginal density

$$\bar{p}_k(x^n) = \int_{\theta \in \Theta_k} p_\theta(x^n) w(\theta) \, \mathrm{d}\theta. \tag{5.1}$$

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, $w(\theta|x^n) = p_\theta(x^n)w(\theta)/\int p_\theta(x^n)w(\theta)\,\mathrm{d}\theta$, 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. We may usually think of the estimators $\bar{p}_k$ as universal codes relative to $\mathcal{M}_k$ [39]. This unified view is known as the *prequential approach to statistics* or *predictive MDL* [27, 69].

Example 14. 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 $X_1, X_2, \ldots$ 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. It is well-known that 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!

In general, for a k-dimensional parametric model $\mathcal{M}_k$, we can define $\bar{p}_k(X_{n+1} \mid x^n) = p_{\hat{\theta}'_k(x^n)}(X_{n+1})$ for some smoothed ML estimator $\hat{\theta}'_k$. The joint distribution with density $\bar{p}_k(x^n)$ will then resemble, but in general not be precisely equal to, the Bayes marginal distribution with density $\bar{p}_k(x^n)$ under some prior on $\mathcal{M}_k$ [39].

5.1.3 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

$$\mathbb{S} = \{(t_1, k_1), \ldots, (t_m, k_m) \in (\mathbb{N} \times \mathbb{Z}^+)^m \mid m \in \mathbb{Z}^+, 0 = t_1 < \ldots < t_m\}. \tag{5.2}$$

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:

$$q_{\mathbf{s}}(X_{n+1}|x^n) = \begin{cases} p_{k_1}(X_{n+1}|x^n) & \text{if } n < t_2, \\ p_{k_2}(X_{n+1}|x^n) & \text{if } t_2 \le n < t_3, \\ \vdots & \vdots \\ p_{k_{m-1}}(X_{n+1}|x^n) & \text{if } t_{m-1} \le n < t_m, \\ p_{k_m}(X_{n+1}|x^n) & \text{if } t_m \le n. \end{cases} \tag{5.3}$$

Switching to the same predictor multiple times is allowed. The extra switch-point t_1 is included to simplify notation; we always take $t_1 = 0$, so that k_1 represents the strategy that is used in the beginning, before any actual switch takes place. Using (5.3), we may now define the switch-distribution as a Bayesian mixture of the elements of $\mathcal{Q}$ according to a prior π on $\mathbb{S}$:

Definition 5.1.1 (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

$$p_{\mathrm{sw}}(x^n) = \sum_{\mathbf{s}\in\mathbb{S}} q_{\mathbf{s}}(x^n) \cdot \pi(\mathbf{s}). \tag{5.4}$$

Although the switch-distribution provides a general way to combine prediction strategies (see Section 5.6), in this paper it will only be applied to combine prediction strategies $\bar{p}_1$, $\bar{p}_2$, ... that correspond to parametric 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 $t_i(\mathbf{s}) \leq n$ and $i = m(\mathbf{s}) \vee n < t_{i+1}(\mathbf{s})$. Now note that by Bayes' theorem, the prior π, together with the data x^n, induces a posterior $\pi(\mathbf{s} \mid x^n) \propto q_{\mathbf{s}}(x^n)\pi(\mathbf{s})$ on switching strategies $\mathbf{s}$. This posterior on switching strategies further induces a posterior on the model K_{n+1} that is used to predict X_{n+1}. Algorithm 5.1, given in Section 5.5, efficiently computes the posterior distribution on K_{n+1} given x^n:

$$\pi(K_{n+1} = k \mid x^n) = \frac{\sum_{\{\mathbf{s}:K_{n+1}(\mathbf{s})=k\}} \pi\big(\mathbf{s}\big) q_{\mathbf{s}}(x^n)}{p_{\mathrm{sw}}(x^n)}, \tag{5.5}$$

which is defined whenever $p_{\mathrm{sw}}(x^n)$ is non-zero. We turn this into a model selection criterion

$$\delta_{\mathrm{sw}}(x^n) = \arg\max_k \pi(K_{n+1} = k | x^n)$$

that selects the model with maximum posterior probability.

5.2 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 5.2.1 states that, if the prediction strategies $\bar{p}_k$ associated with the models are Bayesian predictive distributions, then δ_{sw} is consistent under certain conditions which are only slightly stronger than those required for standard Bayes factor model selection consistency. Theorem 5.2.2 extends the result to the situation where the $\bar{p}_k$ are not necessarily Bayesian.

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 $\bar{P}_{k'}(A) = 0$ [4]. For example, this can usually be shown to hold if (a) the models are nested and (b) for each k, Θ_k is a subset of Θ_{k+1} of w_{k+1}-measure 0. In most interesting applications in which (a) holds, (b) also holds [39]. For consistency of δ_{sw}, we need to strengthen the mutual singularity-condition to a "conditional" mutual singularity-condition: we require 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 independent and identically distributed (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 $\bar{p}_1$, $\bar{p}_2$, ... be Bayesian prediction strategies with respective parameter spaces Θ_1, Θ_2, ... and priors w_1, w_2, ..., and let π be the prior of the corresponding switch-distribution.

Theorem 5.2.1 (Consistency of the Switch-Distribution)**.** *Suppose π is positive everywhere on $\{\mathbf{s} \in \mathbb{S} \mid m(\mathbf{s}) = 1\}$ and such that for some positive constant c, 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*

$$\pi(K_{n+1} = k^* \mid X^n) \stackrel{n\to\infty}{\longrightarrow} 1 \qquad \textit{with } P_{\theta^*}\textit{-probability } 1. \tag{5.6}$$

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

$$\pi(\mathbf{s}) = \pi_{\text{M}}(m)\pi_{\text{K}}(k_1)\prod_{i=2}^{m}\pi_{\text{T}}(t_i|t_i > t_{i-1})\pi_{\text{K}}(k_i), \tag{5.7}$$

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 $c = \theta/(1-\theta)$.

We now extend the theorem to the case where the universal distributions $\bar{p}_1, \bar{p}_2, \ldots$ are not necessarily Bayesian, i.e. they are not necessarily of the form (5.1). It turns out that the "meta-Bayesian" universal distribution P_{sw} is still consistent, as long as the following condition holds. The condition essentially expresses that, for each k, $\bar{p}_k$ must not be too different from a Bayesian predictive distribution based on (5.1). This can be verified if all models $\mathcal{M}_k$ are exponential families (as in, for example, linear regression problems), and the $\bar{p}_k$ represent ML or smoothed ML estimators (see Theorem 2.1 and 2.2 of [57]). We suspect that it holds as well for more general parametric models and universal codes, but we do not know of any proof.

Condition There exist Bayesian prediction strategies $\bar{p}_1^{\text{B}}, \bar{p}_2^{\text{B}}, \ldots$ of form (5.1), with continuous and strictly positive priors $w_1, w_2, \ldots$ such that

1. The conditions of Theorem 5.2.1 hold for $\bar{p}_1^{\text{B}}, \bar{p}_2^{\text{B}}, \ldots$ and the chosen switch-distribution prior π.

2. For all $k \in \mathbb{N}$, for each compact subset Θ' of the interior of Θ_k, there exists a K such that for all $\theta \in \Theta'$, with θ-probability 1, for all n

$$-\log \bar{p}_k(X^n) + \log \bar{p}_k^{\mathrm{B}}(X^n) \leq K.$$

3. For all $k, k' \in \mathbb{N}$ with $k \neq k'$, $\bar{p}_k^{\mathrm{B}}$ and $\bar{p}_{k'}$ are mutually singular.

Theorem 5.2.2 (Consistency of the Switch-Distribution, Part 2)**.** *Let $\bar{p}_1, \bar{p}_2, \ldots$ be prediction strategies and let π be the prior of the corresponding switch distribution. Suppose that the condition above holds relative to $\bar{p}_1, \bar{p}_2, \ldots$ and π. Then, for all $k^* \in \mathbb{N}$, for all $\theta^* \in \Theta_{k^*}$ except for a subset of Θ_{k^*} of Lebesgue-measure 0, the posterior distribution on K_{n+1} satisfies*

$$\pi(K_{n+1} = k^* \mid X^n) \stackrel{n\to\infty}{\longrightarrow} 1 \qquad \textit{with } P_{\theta^*}\textit{-probability } 1. \tag{5.8}$$

5.3 Optimal Risk Convergence Rates

In this section we investigate how well the switch-distribution is able to predict future data in terms of its accumulated KL-risk, which will be formally defined shortly. We first compare predictive performance of the switch-distribution to that achieved by Bayesian model averaging in Section 5.3.1, showing that, reassuringly, the summed risk achieved by P_{sw} is never more than a small constant higher than that achieved by P_{bma}. Then in Section 5.3.2 we describe the general setup and establish a lemma that is used as a general tool in the analysis. Section 5.3.3 treats the case where the data are sampled from a density p^* which is an element of one of the considered models $\mathcal{M}_{k^*}$ for some $k^* \in \mathbb{Z}^+$. In this case we already know from the previous section that the switch-distribution is typically consistent; here we show that it will also avoid the catch-up phenomenon as described in the introduction. Then in Section 5.3.4, we look at the situation where p^* is not in any of the considered models. For this harder, nonparametric case we compare the loss of the switch distribution to that of any other model selection criterion, showing that under some conditions that depend on this reference criterion, the two losses are of the same order of magnitude. Finally in Section 5.3.5 we apply our results to the problem of histogram density estimation, showing that for the class of densities that are (uniformly) bounded away from zero and infinity, and have bounded first derivatives, the switch-distribution (based on histogram models with Bayesian estimators that have uniform priors) predicts essentially as well as any other procedure whatsoever.

The setup is as follows. Suppose $X_1, X_2, \ldots$ are distributed according to a distribution P^* with density $p^* \in \mathcal{M}^*$, where $\mathcal{M}^*$ is an arbitrary set of densities on $\mathcal{X}^\infty$. Specifically, $X_1, X_2, \ldots$ do not have to be i.i.d. We abbreviate "P^* described by density $p^* \in \mathcal{M}^*$" to "$P^* \in \mathcal{M}^*$".

For prediction we use a sequence of parametric models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ with associated estimators $\bar{P}_1, \bar{P}_2, \ldots$ as before. We write $\mathcal{M} = \cup_{i=1}^\infty \mathcal{M}_i$. In Section 5.3.3

we assume that $P^* \in \mathcal{M}$, while in Section 5.3.4 we assume that this is not the case, i.e. $P^* \in \mathcal{M}^* \setminus \mathcal{M}$.

Given $X^{n-1} = x^{n-1}$, we will measure how well any estimator $\bar{P}$ predicts X_n in terms of the Kullback-Leibler (KL) divergence $D(P^*(X_n = \cdot \mid x^{n-1}) \| \bar{P}(X_n = \cdot \mid x^{n-1}))$ [6]. Suppose that P and Q are distributions for some random variable Y, with densities p and q respectively. Then the KL divergence from P to Q is

$$D(P\|Q) = E_P\left[\log \frac{p(Y)}{q(Y)}\right]. \tag{5.9}$$

KL divergence is never negative, and reaches zero if and only if P equals Q. Taking an expectation over X^{n-1} leads to the following (standard) definition of the *risk* of estimator $\bar{P}$ at sample size n relative to KL divergence:

$$R_n(P^*, \bar{P}) = \underset{X^{n-1}\sim P^*}{E}\left[D\big(P^*(X_n = \cdot \mid X^{n-1}) \| \bar{P}(X_n = \cdot \mid X^{n-1})\big)\right]. \tag{5.10}$$

The following identity connects accumulated statistical KL-risk to the information-theoretic redundancy (see e.g. [6] or [39, Chapter 15]) : for all n we have

$$\sum_{i=1}^{n} R_i(P^*, \bar{P}) = \sum_{i=1}^{n} E\left[\log \frac{p^*(X_i \mid X^{i-1})}{\bar{p}(X_i \mid X^{i-1})}\right] = E\left[\log \frac{p^*(X^n)}{\bar{p}(X^n)}\right] = D\left(P^{*(n)} \| \bar{P}^{(n)}\right), \tag{5.11}$$

where the superscript (n) denotes taking the marginal of the distribution on the first n outcomes.

5.3.1 The Switch-distribution vs Bayesian Model Averaging

Here we show that the summed risk achieved by switch-distribution is never much higher than that of Bayesian model averaging, which is itself never much higher than that of any of the estimators $\bar{P}_k$ under consideration.

Lemma 5.3.1. *Let P_{sw} be the switch-distribution for $\bar{P}_1, \bar{P}_2, \ldots$ with prior π of the form (5.7). Let P_{bma} be the Bayesian model averaging distribution for the same estimators, defined with respect to the same prior on the estimators π_{K}. Then, for all $n \in \mathbb{Z}^+$, all $x^n \in \mathcal{X}^n$, and all $k \in \mathbb{Z}^+$,*

$$p_{\mathrm{sw}}(x^n) \quad \geq \quad \pi_{\mathrm{M}}(1) p_{\mathrm{bma}}(x^n) \quad \geq \quad \pi_{\mathrm{M}}(1)\pi_{\mathrm{K}}(k)\bar{p}_k(x^n).$$

Consequently, for all $P^* \in \mathcal{M}^*$ we have

$$\begin{aligned} &\sum_{i=1}^{n} R_i(P^*, P_{\mathrm{sw}}) \\ \leq &\sum_{i=1}^{n} R_i(P^*, P_{\mathrm{bma}}) - \log \pi_{\mathrm{M}}(1) \\ \leq &\sum_{i=1}^{n} R_i(P^*, \bar{P}_k) \quad - \log \pi_{\mathrm{M}}(1) - \log \pi_{\mathrm{K}}(k). \end{aligned}$$

Proof. For the first part we underestimate sums:

$$\begin{aligned} p_{\mathrm{sw}}(x^n) &= \sum_{m \in \mathbb{Z}^+} \sum_{\mathbf{s} \in \mathbb{S}: m(\mathbf{s})=m} q_{\mathbf{s}}(x^n)\pi(\mathbf{s}) \;\geq\; \pi_{\mathrm{M}}(1) \cdot \sum_{k' \in \mathbb{Z}^+} \pi_{\mathrm{K}}(k')\bar{p}_{k'}(x^n) \\ &= \pi_{\mathrm{M}}(1) \cdot p_{\mathrm{bma}}(x^n), \\ p_{\mathrm{bma}}(x^n) &= \sum_{k' \in \mathbb{Z}^+} \bar{p}_{k'}(x^n)\pi_{\mathrm{K}}(k') \;\geq\; \pi_{\mathrm{K}}(k)\bar{p}_k(x^n). \end{aligned}$$

We apply (5.11) to obtain the difference in summed risk:

$$\begin{aligned} \sum_{i=1}^{n} R_i(P^*, P_{\mathrm{sw}}) &= E\left[\log \frac{p^*(X^n)}{p_{\mathrm{sw}}(X^n)}\right] \\ &\leq E\left[\log \frac{p^*(X^n)}{\pi_{\mathrm{M}}(1)p_{\mathrm{bma}}(X^n)}\right] = \sum_{i=1}^{n} R_i(P^*, P_{\mathrm{bma}}) - \log \pi_{\mathrm{M}}(1), \\ \sum_{i=1}^{n} R_i(P^*, P_{\mathrm{bma}}) &= E\left[\log \frac{p^*(X^n)}{p_{\mathrm{bma}}(X^n)}\right] \\ &\leq E\left[\log \frac{p^*(X^n)}{\pi_{\mathrm{K}}(k)\bar{p}_k(X^n)}\right] = \sum_{i=1}^{n} R_i(P^*, \bar{P}_k) - \log \pi_{\mathrm{K}}(k). \end{aligned}$$

□

As mentioned in the introduction, one advantage of model averaging using p_{bma} is that it always predicts almost as well as the estimator $\bar{p}_k$ for *any* k, including the $\bar{p}_k$ that yields the best predictions overall. Lemma 5.3.1 shows that this property is shared by p_{sw}, which multiplicatively dominates p_{bma}. In the following sections, we will investigate under which circumstances the switch-distribution may achieve a *lower* summed risk than Bayesian model averaging.

5.3.2 Improved Convergence Rate: Preliminaries

Throughout our analysis of the achieved rate of convergence we will require that the prior of the switch-distribution, π, can be factored as in (5.7), and is chosen to satisfy

$$-\log \pi_{\mathrm{M}}(m) = O(m), \quad -\log \pi_{\mathrm{K}}(k) = O(\log k), \quad -\log \pi_{\mathrm{T}}(t) = O(\log t). \tag{5.12}$$

Thus π_{M}, the prior on the total number of distinct predictors, is allowed to decrease either exponentially (as required for Theorem 5.2.1) or polynomially, but π_{T} and π_{K} cannot decrease faster than polynomially. For example, we could set $\pi_{\text{T}}(t) = 1/(t(t+1))$ and $\pi_{\text{K}}(k) = 1/(k(k+1))$, or we could take the universal prior on the integers [68].

As competitors to the switch-distribution we introduce a slight generalisation of model selection criteria:

Definition 5.3.2 (Oracle)**.** An *oracle* is a function $\sigma : \mathcal{M}^* \times \mathcal{X}^* \to \mathbb{Z}^+$ that is given not only the observed data $x^n \in \mathcal{X}^*$, but also the generating distribution $P^* \in \mathcal{M}^*$, which it may use to choose a model index $\sigma(P^*, x^n) \in \mathbb{Z}^+$ for all $n \in \mathbb{Z}^+$.

Given an oracle σ, for any P^* and n, x^{n-1}, we abbreviate $\sigma_i = \sigma(P^*, x^{i-1})$ for $1 \leq i \leq n$. We define P_σ as the distribution on X^∞ with marginal densities $p_\sigma(x^n) = \prod_{i=1}^n p_{\sigma_i}(x_i|x^{i-1})$ for all n, x^n. Furthermore, we may split the sequence $\sigma_1, \ldots, \sigma_n$ into segments where the same model is chosen. Now let $m_n(\sigma)$ be the maximum number of such distinct segments over all P^* and all $x^{n-1} \in \mathcal{X}^{n-1}$. That is, let

$$m_n(\sigma) = \max_{P^*} \max_{x^{n-1} \in \mathcal{X}^{n-1}} |\{1 \leq i \leq n-1 : \sigma_i \neq \sigma_{i+1}\}| + 1. \tag{5.13}$$

(The maximum always exists, because for any P^* and x^{n-1} the number of segments is at most n.)

The following lemma expresses that any oracle σ that does not select overly complex models, can be approximated by the switch-distribution with a maximum overhead that depends on $m_n(\sigma)$, its maximum number of segments. We will typically be interested in oracles σ such that this maximum is small in comparison to the sample size, n. The lemma is a tool in establishing the convergence rate of P_{sw}, both in the parametric and the nonparametric contexts considered below.

Lemma 5.3.3. *Let P_{sw} be the switch-distribution, defined with respect to a sequence of estimators $\bar{P}_1, \bar{P}_2, \ldots$ as introduced above, with any prior π that satisfies the conditions in* (5.12) *and let $P^* \in \mathcal{M}^*$. Suppose τ is a positive real number and σ is an oracle such that*

$$\sigma(P^*, x^{i-1}) \leq i^\tau \tag{5.14}$$

for all $i \in \mathbb{Z}^+$, all $x^{i-1} \in \mathcal{X}^{i-1}$. Then

$$\sum_{i=1}^n R_i(P^*, P_{\text{sw}}) = \sum_{i=1}^n R_i(P^*, P_\sigma) + m_n(\sigma) \cdot O(\log n), \tag{5.15}$$

where the multiplicative constant in the big-O notation depends only on τ and the constants implicit in (5.12)*.*

Proof. Using (5.11) we can rewrite (5.15) into the equivalent claim

$$E\left[\log \frac{p_\sigma(X^n)}{p_{\text{sw}}(X^n)}\right] = m_n(\sigma) \cdot O(\log n), \tag{5.16}$$

which we proceed to prove. For all n, $x^n \in \mathcal{X}^n$, there exists a $\mathbf{s} \in \mathbb{S}$ with $m(\mathbf{s}) \leq m_n(\sigma)$ that represents the same sequence of models as σ, so that $q_{\mathbf{s}}(x^i \mid x^{i-1}) = p_{\sigma_i}(x^i \mid x^{i-1})$ for $1 \leq i \leq n$. Consequently, we can bound

$$p_{\text{sw}}(x^n) \;=\; \sum_{\mathbf{s}' \in \mathbb{S}} q_{\mathbf{s}'}(x^n) \cdot \pi(\mathbf{s}') \;\geq\; q_{\mathbf{s}}(x^n)\pi(\mathbf{s}) \;=\; p_\sigma(x^n)\pi(\mathbf{s}). \tag{5.17}$$

By assumption (5.14) we have that σ, and therefore $\mathbf{s}$, never selects a model $\mathcal{M}_k$ with index k larger than i^τ to predict the ith outcome. Together with (5.12) this implies that

$$\begin{aligned}
&-\log \pi(\mathbf{s}) \\
&= -\log \pi_{\text{M}}(m(\mathbf{s})) - \log \pi_{\text{K}}(k_1(\mathbf{s})) + \sum_{j=2}^{m(\mathbf{s})} \big(-\log \pi_{\text{T}}(t_j(\mathbf{s}) \mid t_{j-1}(\mathbf{s})) - \log \pi_{\text{K}}(k_j(\mathbf{s}))\big) \\
&= O(m(\mathbf{s})) + \sum_{j=1}^{m(\mathbf{s})} O(\log t_j(\mathbf{s})) + O(\log k_j(\mathbf{s})) \\
&= O(m(\mathbf{s})) + \sum_{j=1}^{m(\mathbf{s})} O(\log t_j(\mathbf{s})) + O\left(\log\left((t_j(\mathbf{s})+1)^\tau\right)\right) = m_n(\sigma) \cdot O(\log n),
\end{aligned} \tag{5.18}$$

where the multiplicative constant in the big-O in the final expression depends only on τ and the multiplicative constants in (5.12). Together (5.17) and (5.18) imply (5.16), which was to be shown. □

In the following subsections, we compare the accumulated risk of P_{sw} to that of P_σ for various oracles σ; in all these cases our results are independent of the data generating distribution $P^* \in \mathcal{M}^*$. For that reason it will be convenient to define the worst-case summed risk of the switch-distribution and of oracle σ:

$$G_{\text{sw}}(n) := \sup_{P^* \in \mathcal{M}^*} \sum_{i=1}^{n} R_i(P^*, P_{\text{sw}}), \text{ and} \tag{5.19}$$

$$G_\sigma(n) := \sup_{P^* \in \mathcal{M}^*} \sum_{i=1}^{n} R_i(P^*, P_\sigma). \tag{5.20}$$

We will also compare the accumulated risk of P_{sw} to the *minimax risk in sum*, defined as

$$G_{\text{mm-fix}}(n) := \inf_{\bar{P}} \sup_{P^* \in \mathcal{M}^*} \sum_{i=1}^{n} R_i(P^*, \bar{P}). \tag{5.21}$$

Here the infimum is over all prequential forecasting systems $\bar{P}$ for which, for each n, $x^{n-1} \in \mathcal{X}^{n-1}$, $\bar{P}(X_n = \cdot \mid X^{n-1} = x^{n-1})$ admits a density. Equivalently, the infimum is over all sequences of n estimators $\bar{P}(X_1), \bar{P}(X_2 \mid X^1), \ldots, \bar{P}(X_n \mid X^{n-1})$. Note that there is no requirement that $\bar{P}$ maps x^n to a distribution in $\mathcal{M}^*$ or $\mathcal{M}$; we are looking at the worst case over all possible estimators, irrespective of the model $\mathcal{M}$ used to approximate $\mathcal{M}^*$. Thus, we may call $\bar{P}$ an "out-model estimator" [39]. The notation $G_{\text{mm-fix}}$ will be clarified in Section 5.4, where we compare convergence in sum with more standard notions of convergence.

5.3.3 The Parametric Case

Here we assume that $P^* \in \mathcal{M}_{k^*}$ for some $k^* \in \mathbb{Z}^+$, but we also consider that if $\mathcal{M}_1, \mathcal{M}_2, \ldots$ are of increasing complexity, then the catch-up phenomenon may occur, meaning that at small sample sizes, some estimator $\bar{P}_k$ with $k < k^*$ may achieve lower risk than $\bar{P}_{k^*}$. The following lemma shows that the predictive performance of the switch-distribution is never much higher than the predictive performance of the best oracle that iterates through the models in order of increasing complexity.

Lemma 5.3.4. *Let P_{sw} be the switch distribution, defined with respect to a sequence of estimators $\bar{P}_1, \bar{P}_2, \ldots$ as above, with prior π satisfying* (5.12)*. Let $k^* \in \mathbb{Z}^+$, and let σ be any oracle such that for all P^*, all x^∞, we have that $\sigma(P^*, x^n)$ is monotonically nondecreasing in n; furthermore for sufficiently large n, we have $\sigma(P^*, x^n) = k^*$. Then*

$$G_{\text{sw}}(n) - G_\sigma(n) \leq \sup_{P^* \in \mathcal{M}^*} \left(\sum_{i=1}^{n} R_i(P^*, P_{\text{sw}}) - \sum_{i=1}^{n} R_i(P^*, P_\sigma) \right) = k^* \cdot O(\log n).$$

In particular, if for some $c' > 0$, for all sufficiently large n, $G_\sigma(n) \geq c \log n$ (i.e. $G_\sigma = \Omega(\log n)$), then there is a c such that

$$\limsup_{n \to \infty} \frac{G_{\text{sw}}(n)}{G_\sigma(n)} \leq c.$$

The additional risk compared to P_σ is of order $\log n$. In the parametric case, we often have $G_{\text{mm-fix}}(n)$ proportional to $\log n$ (Section 5.4.1). If that is the case, and if, as seems reasonable, there is an oracle σ that satisfies the given restrictions and that achieves summed risk proportional to $G_{\text{mm-fix}}(n)$, then also the switch-distribution achieves a proportional summed risk.

Proof. The first inequality is a consequence of the general rule that for two functions f and g, we have $\sup_x f(x) - \sup_x g(x) \leq \sup_x (f(x) - g(x))$. We proceed

to show the asymptotics of the second term, which has the supremum on the outside. Let $\mathbf{s} = (0, k^*)$. We have, uniformly for all n, $x^n \in \mathcal{X}^n$,

$$-\log p_{\mathrm{sw}}(x^n) \;=\; -\log \sum_{\mathbf{s}' \in \mathbb{S}} q_{\mathbf{s}'}(x^n)\pi(\mathbf{s}') \;\leq\; -\log q_{\mathbf{s}}(x^n) + \pi(\mathbf{s}). \tag{5.22}$$

Since σ satisfies (5.14) for suitably chosen τ, and the properties of σ ensure $m_n(\sigma) \leq k^*$, we can apply Lemma 5.3.3. The first part of the lemma is then obtained by taking the supremum over P^*. To establish the second part, we can choose a c such that

$$\limsup_{n\to\infty} \frac{G_{\mathrm{sw}}(n)}{G_\sigma(n)} \leq \limsup_{n\to\infty} \frac{G_\sigma(n) + k^* \cdot O(\log n)}{G_\sigma(n)} \leq 1 + \limsup_{n\to\infty} \frac{k^* \cdot O(\log n)}{c' \log n} = c. \tag{5.23}$$

□

5.3.4 The Nonparametric Case

In this section we develop an analogue of Lemma 5.3.4 for the nonparametric case, where there is no k such that $P^* \in \mathcal{M}_k$. It is then applied to the problem of histogram density estimation.

Lemma 5.3.5. *Let P_{sw} be the switch-distribution, defined with respect to a sequence of estimators $\bar{P}_1, \bar{P}_2, \ldots$ as above, with any prior π that satisfies the conditions in (5.12). Let $f : \mathbb{Z}^+ \to [0, \infty)$ and let $\mathcal{M}^*$ be a set of distributions on X^∞. Suppose there exist an oracle σ and positive constants τ and c such that*

(i) $\sigma(P^*, x^{i-1}) \leq i^\tau$ *for all* $i \in \mathbb{Z}^+$, *all* $x^{i-1} \in \mathcal{X}^{i-1}$,

(ii) $m_n(\sigma) \log n = o(f(n))$, *and*

(iii) $\limsup_{n\to\infty} \frac{G_\sigma(n)}{f(n)} \leq c$.

Then

$$\limsup_{n\to\infty} \frac{G_{\mathrm{sw}}(n)}{f(n)} \leq c. \tag{5.24}$$

Proof. By (i) we can apply Lemma 5.3.3 to σ. Using conditions (ii) and (iii), a derivation similar to (5.23) completes the proof:

$$\limsup_{n\to\infty} \frac{G_{\mathrm{sw}}(n)}{f(n)} \leq$$
$$\limsup_{n\to\infty} \frac{G_\sigma(n) + m_n(\sigma)O(\log n)}{f(n)} \leq c + \limsup_{n\to\infty} \frac{m_n(\sigma)O(\log n)}{f(n)} = c. \qquad \square$$

Lemma 5.3.5 can be used to show minimax rates of convergence relative to specific nonparametric model classes $\mathcal{M}^*$. The general idea is to apply the lemma with $f(n)$ equal to the minimax risk in sum $G_{\text{mm-fix}}(n)$ (see (5.21)). It will be seen that in many standard nonparametric settings, one can exhibit an oracle σ that only switches sporadically (Condition (ii) of the lemma) and that achieves $G_{\text{mm-fix}}(n)$ (Condition (iii)). The lemma then implies that P_{sw} achieves the minimax risk as well. As a proof of concept, we now show this in detail for histogram density estimation. In Section 5.4.2, we discuss possibilities for extending the reasoning to more general settings.

5.3.5 Example: Histogram Density Estimation

Rissanen, Speed and Yu [74] consider density estimation based on histogram models with equal-width bins relative to a restricted set $\mathcal{M}^*$ of "true" distributions, identified in this section by their densities on the unit interval $\mathcal{X} = [0, 1]$. The restriction on $\mathcal{M}^*$ is that there should exist constants $0 < c_0 < 1 < c_1$ such that for every density $p \in \mathcal{M}^*$, for all $x \in \mathcal{X}$, $c_0 \leq p(x) \leq c_1$ and $|p'(x)| \leq c$, where p' denotes the first derivative of p; unlike in the paper we require a uniform bound c on the derivative which may not depend on $p \in \mathcal{M}^*$. The densities are extended to sequences by independence: $p(x^n) \equiv p^n(x^n) = \prod_{i=1}^n p(x_i)$ for $x^n \in \mathcal{X}^n$.

The histogram model $\mathcal{M}_k$ is the set of densities on $\mathcal{X} = [0, 1]$ that are constant within the k bins $[0, a_1]$, $(a_1, a_2]$, ..., $(a_{k-1}, 1]$, where $a_i = i/k$, i.e. $\mathcal{M}_k$ contains all densities with p_θ such that, for all $x, x' \in [0, 1]$, if x and x' lie in the same bin, then $p_\theta(x) = p_\theta(x')$. The $k-1$-dimensional vector $\theta = (\theta_1, \ldots, \theta_{k-1})$ denotes the probability masses of the first $k-1$ bins. The last bin then gets the remaining mass, $1 - \sum_{j=1}^{k-1} \theta_k$. Note that the number of free parameters is one less than the number of bins.

Here we model densities from $\mathcal{M}^*$ by sequences of densities based on histogram models of an increasing number of bins as more data become available. Rissanen, Speed and Yu define prediction strategies $\bar{p}_k$ on $\mathcal{X}^\infty$ relative to each histogram model $\mathcal{M}_k$. For model $\mathcal{M}_k$, the conditional predictions are given by

$$\bar{p}_k(x_{n+1} \mid x^n) := \frac{n_{x_{n+1}}(x^n) + 1}{n + k} \cdot k, \tag{5.25}$$

where $n_{x_{n+1}}(x^n)$ denotes the number of outcomes in x^n that fall into the same bin as x_{n+1}. These $\bar{p}_k$ correspond to Bayesian estimators relative to a uniform prior on the set of parameters, $\{\theta\}$.

Yu [107] shows that, relative to the $\mathcal{M}^*$ defined above, the minimax-risk in sum satisfies

$$\limsup_{n \to \infty} \frac{G_{\text{mm-fix}}(n)}{n^{1/3}} = C,$$

where C is a constant that depends on the constants c, c_0, c_1 used in the definition of $\mathcal{M}^*$. In [74] it is shown that the simple strategy that uses the histogram model

with $\left\lceil n^{1/3} \right\rceil$ bins to predict the nth outcome achieves this minimax risk in sum up to a constant multiplicative factor:

Theorem 5.3.6 (Theorem 1 from [74]). *For all $p^* \in \mathcal{M}^*$*

$$\limsup_{n\to\infty} \frac{\sum_{i=1}^{n} R_i(P^*, \bar{P}_{\lceil n^{1/3} \rceil})}{n^{1/3}} \leq A_{p^*}, \tag{5.26}$$

where A_{p^*} is a constant that depends only on c_{p^*}, the bound on the first derivative of p^*.

We will now show that the switch-distribution also achieves the minimax risk in sum up to the same constant factor. The idea is to view $\bar{P}_{\lceil n^{1/3} \rceil}$ as an oracle. Even though it makes no use of P^* in its selection, $\bar{P}_{\lceil n^{1/3} \rceil}$ clearly satisfies Definition 5.3.2, the definition of an oracle. We would like to apply Lemma 5.3.5 to oracle $\bar{P}_{\lceil n^{1/3} \rceil}$, but we cannot do so, since the oracle switches prediction strategies polynomially often in n. To prove that P_{sw} achieves a rate of $n^{1/3}$, we first need to consider a cruder version of $\bar{P}_{\lceil n^{1/3} \rceil}$ that still achieves the minimax rate, but only switches logarithmically often. This is the key to the proof of Theorem 5.3.7.

Theorem 5.3.7. *Let p_{sw} denote the switch-distribution with prior π that satisfies the conditions in* (5.12)*, relative to histogram prediction strategies $\bar{p}_1$, $\bar{p}_2$, ... For all $p^* \in \mathcal{M}^*$ this switch-distribution satisfies*

$$\limsup_{n\to\infty} \frac{\sum_{i=1}^{n} R_i(P^*, P_{\text{sw}})}{n^{1/3}} \leq A_{p^*}, \tag{5.27}$$

where A_{p^*} is the same constant as in (5.26)*.*

We first note that in [74], Theorem 5.3.6 is proved from the following more general theorem, which gives an upper bound on the risk of any prediction strategy that uses a histogram model with approximately $\left\lceil n^{1/3} \right\rceil$ bins to predict outcome n:

Theorem 5.3.8. *For any $\alpha \geq 1$, any $k \in [\left\lceil (n/\alpha)^{1/3} \right\rceil, \left\lceil n^{1/3} \right\rceil]$, and any $p^* \in \mathcal{M}^*$,*

$$R_n(P^*, \bar{P}_k) \leq \alpha^{2/3} C_{p^*} n^{-2/3}, \tag{5.28}$$

where C_{p^*} depends only on the upper bound c on the first derivative of p^*.

In [74] the theorem is only proven for $\alpha = 1$, but the proof remains valid unchanged for any $\alpha > 1$. From this, Theorem 5.3.6 follows by summing (5.28), and approximating $\sum_{i=1}^{n} i^{-2/3}$ by an integral. We will now apply it to prove Theorem 5.3.7 as well.

Proof of Theorem 5.3.7. Choose any $p^* \in \mathcal{M}^*$, and let $\alpha > 1$ be arbitrary. Let $t_j = \lceil \alpha^{j-1} \rceil - 1$ for $j \in \mathbb{Z}^+$ be a sequence of switch-points, and define an oracle $\sigma_\alpha(P^*, x^{n-1}) := \lceil (t_j+1)^{1/3} \rceil$ for any $x^{n-1} \in \mathcal{X}^{n-1}$, where j is chosen such that $n \in [t_j + 1, t_{j+1}]$. By applying Lemma 5.3.5 to σ_α with $f(n) = n^{1/3}$, $\tau = 1$ and $c = \alpha^{2/3} A_{p^*}$, we immediately obtain

$$\limsup_{n\to\infty} \frac{\sum_{i=1}^n R_i(P^*, P_{\mathrm{sw}})}{n^{1/3}} \leq \alpha^{2/3} A_{p^*} \tag{5.29}$$

for any $\alpha > 1$. Theorem 5.3.7 then follows, because the left-hand side of this expression does not depend on α. We still need to show that conditions (i)–(iii) of Lemma 5.3.5 are satisfied.

Note that σ_α uses approximately $\left\lceil n^{1/3} \right\rceil$ bins to predict the nth outcome, but has relatively few switch-points: it satisfies $m_n(\sigma_\alpha) \leq \lceil \log_\alpha n \rceil + 2$. Thus, conditions (i) and (ii) are comfortably satisfied. To verify (iii), note that the selected number of bins is close to $\left\lceil n^{1/3} \right\rceil$ in the sense of Theorem 5.3.8: For $n \in [t_j + 1, t_{j+1}]$ we have

$$\left\lceil (t_j+1)^{1/3} \right\rceil = \left\lceil \left(\frac{n}{n/(t_j+1)} \right)^{1/3} \right\rceil \in \left[\left\lceil (n/\alpha)^{1/3} \right\rceil, \left\lceil n^{1/3} \right\rceil \right] \tag{5.30}$$

using $n \leq t_{j+1}$ and $(t_{j+1})/(t_j+1) \leq \alpha$. We can now apply Theorem 5.3.8 to obtain

$$\limsup_{n\to\infty} \frac{\sum_{i=1}^n R_i(P^*, \sigma_\alpha)}{n^{1/3}} \leq \limsup_{n\to\infty} \frac{\sum_{i=1}^n \alpha^{2/3} C_{p^*} i^{-2/3}}{n^{1/3}} \leq \alpha^{2/3} A_{p^*}, \tag{5.31}$$

showing that (iii) is satisfied and Lemma 5.3.5 can be applied to prove the theorem. □

Theorem 5.3.7 shows that the switch distribution obtains the optimal convergence rate in sum relative to $\mathcal{M}^*$. In [74] it is also shown that standard two-part MDL does *not* achieve this rate; it is slower by an $O(\log n)$ factor. The analysis leading to this result also strongly suggests that Bayesian model averaging based on a discrete prior on the Bayesian marginal distributions $\bar{p}_1, \bar{p}_2, \ldots$ given by (5.25) is also a factor $O(\log n)$ slower compared to the minimax rate [39]. By Theorem 5.3.6, $\delta(x^n) := \lceil n^{1/3} \rceil$ defines a very simple model selection criterion which does achieve the optimal rate in sum relative to $\mathcal{M}^*$, but, in contrast to the switch distribution, it is inconsistent. Moreover, if we are lucky enough to be in a scenario where p^* actually allows a *faster* than minimax optimal in-sum convergence by letting the number of bins grow as n^γ for some $\gamma \neq \frac{1}{3}$, the switch-distribution would be able to take advantage of this whereas δ cannot.

5.4 Further Discussion of Convergence in the Nonparametric Case

In Section 5.4.1 we analyse the relation between our convergence rates in sum and standard convergence rates. In Section 5.4.2, we explore possible future applications of Lemma 5.3.5 to establish minimax convergence rates for model classes $\mathcal{M}_1, \mathcal{M}_2, \ldots$ beyond histograms.

5.4.1 Convergence Rates in Sum

Let $g : \mathbb{N} \to \mathbb{R}^+$ and $h : \mathbb{N} \to \mathbb{R}^+$ be two functions that converge to 0 with increasing n. We say that g converges to 0 at rate h if $\limsup_{n\to\infty} \frac{g(n)}{h(n)} \leq 1$. We say that g converges to 0 *in sum* at rate h if $\limsup_{n\to\infty} \frac{\sum_{i=1}^n g(i)}{\sum_{i=1}^n h(i)} \leq 1$. This notion of convergence has been considered by, among others, Rissanen, Speed and Yu [74], Barron [6], Poland and Hutter [65], and was investigated in detail by Grünwald [39]. Note that, if g converges to 0 at rate h, and $\lim_{n\to\infty} \sum_{i=1}^n h(n) = \infty$, then g also converges in sum at rate h. Conversely, suppose that g converges in sum at rate h. Does this also imply that g converges to 0 at rate h in the ordinary sense? The answer is "almost": as shown in [39], $g(n)$ may be strictly greater than $h(n)$ for some n, but the *gap* between any two n and $n' > n$ at which g is larger than h must become infinitely large with increasing n.

We will now informally compare the "convergence in sum" results of the previous section with more standard results about individual risk convergence. We will write $h(n) \asymp g(n)$ if for some $0 < c_1 < c_2$, for all large n, $c_1 g(n) < h(n) < c_2 g(n)$. The minimax convergence rate relative to a set of distributions $\mathcal{M}^*$ is defined as

$$h_{\text{mm}}(n) = \inf_{\bar{P}} \sup_{P^* \in \mathcal{M}} R_n(P^*, \bar{P}), \tag{5.32}$$

where $\bar{P}$ is any estimator that allows a density, it is not required to lie in $\mathcal{M}^*$ or $\mathcal{M}$. If a sequence of estimators achieves (5.32) to within a constant factor, we say that it converges at the "minimax optimal rate." Such a sequence of estimators will also achieve the minimax rate in sum, defined as

$$G_{\text{mm-var}}(n) = \sum_{i=1}^n h_{\text{mm}}(i) = \inf_{\bar{P}} \sum_{i=1}^n \sup_{P^* \in \mathcal{M}^*} R_i(P^*, \bar{P}), \tag{5.33}$$

where $\bar{P}$ now ranges over all prequential forecasting systems (i.e. sequences of estimators). In many nonparametric density estimation and regression problems, the minimax risk $h_{\text{mm}}(n)$ is of order $n^{-\gamma}$ for some $1/2 < \gamma < 1$ (see, for example, [105, 106, 9]), i.e. $h_{\text{mm}}(n) \asymp n^{-\gamma}$, where γ depends on the smoothness assumptions on the densities in $\mathcal{M}^*$. We call the situation with $\mathcal{M}^*$ such that $h_{\text{mm}}(n) \asymp n^{-\gamma}$

the "standard nonparametric case." In this standard case, we have

$$G_{\text{mm-var}}(n) \asymp \sum_{i=1}^{n} i^{-\gamma} \asymp \int_1^n x^{-\gamma}\,\mathrm{d}x \asymp n^{1-\gamma}. \tag{5.34}$$

Similarly, in standard parametric problems, the minimax risk $h_{\text{mm}}(n) \asymp 1/n$. In that case, analogously to (5.34), we see that the minimax risk in sum $G_{\text{mm-var}}$ is of order $\log n$.

Note, however, that our result for histograms (and, more generally, for any rate-of-convergence result that may be based on Lemmata 5.3.3, 5.3.4 and 5.3.5), is based on a scenario where P^*, while allowed to depend on n, is kept fixed over the terms in the sum from 1 to n. Indeed, in Theorem 5.3.7 we showed that P_{sw} achieves the minimax rate in sum $G_{\text{mm-fix}}(n)$ as defined in (5.21). Comparing to (5.33), we see that the supremum is moved outside of the sum. Fortunately, $G_{\text{mm-fix}}$ and $G_{\text{mm-var}}$ are usually of the same order: in the parametric case, e.g. $\mathcal{M}^* = \bigcup_{k \le k^*} \mathcal{M}_k$, both $G_{\text{mm-fix}}$ and $G_{\text{mm-var}}$ are of order $\log n$. For $G_{\text{mm-var}}$, we have already seen this. For $G_{\text{mm-fix}}$, this is a standard information-theoretic result, see for example [23]. In the standard nonparametric case, when the standard minimax rate is of order $n^{-\gamma}$ and therefore $G_{\text{mm-var}} \asymp n^{1-\gamma}$, it also holds that $G_{\text{mm-fix}}(n) \asymp n^{1-\gamma}$ [106, page 1582]. To see this, let $P_{\text{mm-fix}}$ be any prequential forecasting system that achieves $G_{\text{mm-fix}}$ as defined in (5.21) (if such a $P_{\text{mm-fix}}$ does not exist, take any P that, for each n, achieves the infimum to within ε_n for some sequence $\varepsilon_1, \varepsilon_2, \ldots$ tending to 0). Now define the prequential forecasting system

$$P_{\text{Césaro}}(x_n \mid x^{n-1}) := \frac{1}{n} \sum_{i=1}^{n} P_{\text{mm-fix}}(x_n \mid x^{i-1}).$$

Thus, $P_{\text{Césaro}}$ is obtained as a time ("Césaro"-) average of $P_{\text{mm-fix}}$. It now follows by applying Jensen's inequality as in Proposition 15.2 of [39] (or the corresponding results in [102] or [106]) that

$$\sup_{P^*} R_n(P^*, P_{\text{Césaro}}) \le \sup_{P^*} \frac{1}{n} \sum_{i=1}^{n} R_i(P^*, P_{\text{mm-fix}}) = n^{-1} O(n^{1-\gamma}) = O(n^{-\gamma}), \tag{5.35}$$

so that $\sum_{j=1}^{n} \sup_{P^*} R_j(P^*, P_{\text{Césaro}}) = O(\sum_{j=1}^{n} j^{-\gamma}) = O(n^{1-\gamma})$, and it follows that

$$G_{\text{mm-var}}(n) = O(n^{1-\gamma}) = O(G_{\text{mm-fix}}(n)). \tag{5.36}$$

Since, trivially, $G_{\text{mm-fix}}(n) \le G_{\text{mm-var}}(n)$, it follows that $G_{\text{mm-var}}(n) \asymp n^{1-\gamma}$. The underlying idea of basing predictive distributions on Césaro-averages is not new; see for example [43, Section 9] and [102]. It is described in detail in [39].

Summarising, both in standard parametric and nonparametric cases, $G_{\text{mm-fix}}$ and $G_{\text{mm-var}}$ are of comparable size. Therefore, Lemma 5.3.4 and 5.3.5 do suggest

that, both in standard parametric and nonparametric cases, P_{sw} achieves the minimax convergence rate $G_{\text{mm-fix}}$ (and Theorem 5.3.7 shows that it actually does so in a specific nonparametric case). One caveat is in order though: the fact that $G_{\text{mm-fix}}$ and $G_{\text{mm-var}}$ are of comparable size does *not* imply that P_{sw} also achieves the varying-P^*-minimax rate $G_{\text{mm-var}}$. We cannot prove the analogue of Lemma 5.3.4 and Lemma 5.3.5 with the supremum over P^* inside the sum in G_{sw} and G_σ. Therefore, we cannot prove even for histogram density estimation, that P_{sw} achieves $G_{\text{mm-var}}$. Nevertheless, we do suspect that in the standard nonparametric case, $G_{\text{mm-fix}}(n) \asymp G_{\text{mm-var}}(n) \asymp n^{1-\gamma}$, whenever P_{sw} achieves the fixed-P^* minimax rate $G_{\text{mm-fix}}$, it also achieves the varying-P^* minimax rate $G_{\text{mm-var}}$. The reason for this conjecture is that, if we can assume that the data are i.i.d. under all $P^* \in \mathcal{M}^*$, then whenever P_{sw} achieves $G_{\text{mm-fix}}$, a small modification of P_{sw} will achieve $G_{\text{mm-var}}$. Indeed, define the *Césaro-switch distribution* as

$$P_{\text{Césaro-sw}}(x_n \mid x^{n-1}) := \frac{1}{n} \sum_{i=1}^{n} P_{\text{sw}}(x_n \mid x^{i-1}).$$

Applying (5.35) to $P_{\text{Césaro-sw}}$ rather than $P_{\text{Césaro}}$, we see that $P_{\text{Césaro-sw}}$ achieves the varying-P^*-minimax rate whenever P_{sw} achieves the fixed-P^*-minimax rate. Since, intuitively, $P_{\text{Césaro-sw}}$ learns "slower" than P_{sw}, we suspect that P_{sw} itself then achieves the varying-P^*-minimax rate as well. However, in the parametric case, $h_{\text{mm}}(n) \asymp 1/n$ whereas $G_{\text{mm-fix}}(n)/n \asymp (\log n)/n$. Then the reasoning of (5.35) does not apply any more, and $P_{\text{Césaro-sw}}$ may not achieve the minimax rate for varying P^*. We suspect that this is not a coincidence — a recent result by Yang [103] suggests that, in the parametric case, *no* model selection/averaging criterion can achieve both consistency and minimax optimal varying-P^* rates $G_{\text{mm-var}}$ (Section 5.6.1).

5.4.2 Beyond Histograms

The key to proving Theorem 5.3.7, the minimax convergence result for histogram density estimation, is the existence of an oracle σ_α which achieves the minimax convergence rate, but which, at the same time, switches only a logarithmic number of times. The theorem followed by applying Lemma 5.3.5 with this oracle. It appears that the same technique can be applied in many other standard nonparametric settings (with $h_{\text{mm}}(n) \asymp n^{-\gamma}$) as well. Important examples include linear regression based on full approximation sets of functions such as polynomials [106, 101] or splines, with random i.i.d. design and i.i.d. normally distributed noise with known variance σ^2. The development in Section 6.2 of [101] indicates that an analogue of Theorem 5.3.7 can be proven for such cases. Here the models $\mathcal{M}_k$ are families of conditional distributions P_θ for $Y \in \mathbb{R}$ given $X \in [0,1]^d$ for some $d > 0$, where $\theta = (\theta_1, \ldots, \theta_k)$ and P_θ expresses that $Y_i = \sum_{j=1}^{k} \theta_j \phi_j(X_i) + U$, with ϕ_j being the j-th basis function in the approximation set, and U, the noise,

is a zero-mean Gaussian random variable. The forecasting systems $\bar{p}_1, \bar{p}_2, \ldots$ are now based on maximum likelihood estimators rather than Bayes predictive distributions.

Another candidate is density estimation based on sequences of exponential families as introduced by Barron and Sheu [9], when the estimators $\bar{p}_1, \bar{p}_2, \ldots$ are based on Bayesian MAP estimators defined with respect to k-dimensional exponential families, and $\mathcal{M}^*$ is taken to be the set of densities p such that $\log p$ is contained in some *Sobolev space* with smoothness parameter r [7]. Preliminary investigations suggest that $P_{\rm sw}$ achieves the minimax convergence rates in both the linear regression and the density estimation setting, but, at the time of writing, we have not yet proven any formal statements.

5.5 Efficient Computation

For priors π as in (5.7), the posterior probability on predictors $p_1, p_2, \ldots$ can be efficiently computed sequentially, provided that $\pi_{\rm T}(Z = n \mid Z \geq n)$ and $\pi_{\rm K}$ can be calculated quickly and that $\pi_{\rm M}(m) = \theta^m(1-\theta)$ is geometric with parameter θ, as is also required for Theorem 5.2.1 and permitted in the theorems and lemma's of Section 5.3 that require (5.12). The algorithm resembles FIXED-SHARE [46], but whereas FIXED-SHARE implicitly imposes a geometric distribution for $\pi_{\rm T}$, we allow general priors by varying the shared weight with n, and through the addition of the $\pi_{\rm M}$ component of the prior, we ensure that the additional loss compared to the best prediction strategy does not grow with the sample size, a crucial property for consistency.

To ensure finite running time, rather than $P_{\rm sw}$ the algorithm uses a potentially defective distribution P that assigns smaller or equal probability to all events. It is obtained by restricting $P_{\rm sw}$ to using only a finite nonempty set $\mathcal{K}_n$ of prediction strategies at sample size n. Then, analogously to (5.4), for any n the density of the marginal distribution of P on X^n is given by $p(x^n) = \sum_{\mathbf{s} \in \mathbb{S}'} q_{\mathbf{s}}(x^n) \cdot \pi(\mathbf{s})$, where $\mathbb{S}' := \{\mathbf{s} \in \mathbb{S} \mid \forall n \in \mathbb{Z}^+ : K_n(\mathbf{s}) \in \mathcal{K}_n\}$ denotes the parameter space restricted to those prediction strategies that are considered at each sample size. We use the indicator function, $\mathbf{1}_A(x) = 1$ if $x \in A$ and 0 otherwise. Here is the algorithm:

This algorithm can be used to obtain fast convergence in the sense of Section 5.3, and, as long as π does not vary with n, consistency in the sense of Theorem 5.2.1. Note that the running time $\Theta(\sum_{n=1}^{N} |\mathcal{K}_n|)$ is typically of the same order as that of fast model selection criteria like AIC and BIC: for example if the number of considered prediction strategies is fixed at K then the running time is $\Theta(K \cdot N)$. In the interest of clarity and simplicity we only prove the theorem below, which assumes $P = P_{\rm sw}$, but the reader may verify that the algorithm remains valid for defective P.

Theorem 5.5.1. *If $P = P_{\rm sw}$, then Algorithm 5.1 correctly reports $P(K_{n+1}, x^n)$.*

Algorithm 5.1 Switch(x^N)

for $k \in \mathcal{K}_1$ **do** initialise $w_k^a \leftarrow \theta \cdot \pi_{\text{K}}(k)$; $w_k^b \leftarrow (1-\theta) \cdot \pi_{\text{K}}(k)$ **end for**
for $n=1,\ldots,N$ **do**
  Report $P(K_n, x^{n-1}) = w^a_{K_{n+1}} + w^b_{K_{n+1}}$ (a K-sized array)
  for $k \in \mathcal{K}_n$ **do** $w_k^a \leftarrow w_k^a \cdot p_k(x_n|x^{n-1})$; $w_k^b \leftarrow w_k^b \cdot p_k(x_n|x^{n-1})$ **end for**
  $\texttt{pool} \leftarrow \pi_{\text{T}}(Z = n \mid Z \geq n) \cdot \sum_{k\in\mathcal{K}_n} w_k^a$
  for $k \in \mathcal{K}_{n+1}$ **do**
    $w_k^a \leftarrow w_k^a \cdot \mathbf{1}_{\mathcal{K}_n}(k) \cdot \pi_{\text{T}}(Z \neq n \mid Z \geq n) \;+\; \texttt{pool} \cdot \pi_{\text{K}}(k) \cdot \theta$
    $w_k^b \leftarrow w_k^b \cdot \mathbf{1}_{\mathcal{K}_n}(k) \;+\; \texttt{pool} \cdot \pi_{\text{K}}(k) \cdot (1-\theta)$
  end for
end for
Report $P(K_{N+1}, x^N) = w^a_{K_{N+1}} + w^b_{K_{N+1}}$

The proof is given in Section 5.9.4.

5.6 Relevance and Earlier Work

5.6.1 AIC-BIC; Yang's Result

Over the last 25 years or so, the question 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 (googling "AIC *and* BIC" gives 355000 hits) [85, 41, 40, 10, 33, 30, 81]. 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"), one should not expect procedures that combine the best of both types of approaches to exist [81]. Our Theorem 5.2.1 and our results in Section 5.3 show that, at least in some cases, one can get the best of both worlds after all, and model averaging based on P_{sw} achieves the minimax optimal convergence rate. In typical parametric settings ($P^* \in \mathcal{M}$), model selection based on P_{sw} is consistent, and Lemma 5.3.4 suggests that model averaging based on P_{sw} is within a constant factor of the minimax optimal rate in parametric settings. Superficially, our results may seem to contradict the central conclusion of Yang [103]. Yang 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's result is proved for a variation of linear regression in which the estimation error is measured on the previously observed design points. This setup cannot be directly embedded in our framework. Also, Yang's notion of model combination is somewhat different from the model averaging that is used to compute P_{sw}. Thus, formally, there is no contradiction between Yang's results and

ours. Still, the setups are so similar that one can easily imagine a variation of Yang's result to hold in our setting as well. Thus, it is useful to analyse how these "almost" contradictory results may coexist. We suspect (but have no proof) that the underlying reason is the definition of our minimax convergence rate in sum (5.21) in which P^* is allowed to depend on n, but then the risk with respect to that same P^* is summed over all $i = 1, \ldots, n$. Yang's result holds in a parametric scenario, where there are two nested parametric models, and data are sampled from a distribution in one of them. Then both $G_{\text{mm-fix}}$ and $G_{\text{mm-var}}$ are of the same order $\log n$, but it may of course be possible that there does exist a minimax optimal procedure that is also consistent, relative to the $G_{\text{mm-fix}}$-game, in which P^* is kept fixed once n has been determined, while there does not exist a minimax optimal procedure that is also consistent, relative to the $G_{\text{mm-var}}$-game, in which P^* is allowed to vary. Indeed, while in Section 5.4.1 we have established that $P_{\text{Césaro-sw}}$, a slight variation of P_{sw}, achieves the minimax optimal convergence rates $G_{\text{mm-var}}$ and h_{mm} for some nonparametric $\mathcal{M}^*$, which suggests that P_{sw} achieves these rates as well, we do not have such a result for parametric $\mathcal{M}^*$. Yang's results indicate that such an analogue may not exist.

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* [30, 21], and Hansen and Yu's *gMDL* version of MDL regression [41]; also the "mongrel" procedure of [99] 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 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 [104] to select between AIC and BIC in regression problems. Yang shows that a particular form of cross-validation will asymptotically select AIC in case the use of AIC leads to better predictions, and BIC in the case that BIC leads to better predictions. In contrast to Yang, we use a single paradigm rather than a mix of several ones (such as AIC, BIC and cross-validation) – essentially our paradigm is just that of universal individual-sequence prediction, or equivalently, the individual-sequence version of predictive MDL, or equivalently, Dawid's prequential analysis applied to the log scoring rule. Indeed, our work has been heavily inspired by prequential ideas; in Dawid [29] it is already suggested that model selection should be based on the *transient* behaviours in terms of sequential prediction of the estimators within the models: one should select the model which is optimal at the given sample size, and this will change over time. Although Dawid uses standard Bayesian mixtures of parametric models as his running examples, he implicitly suggests that other ways (the details of which are left unspecified) of combining predictive distributions relative to parametric models may be preferable, especially in the nonparametric case where the true distribution is outside any of the parametric models under consideration.

5.6.2 Prediction with Expert Advice

Since the switch-distribution has been designed to perform well in a setting where the optimal predictor $\bar{p}_k$ changes over time, our work is also closely related to the algorithms for *tracking the best expert* in the universal prediction literature [46, 95, 94, 63]. However, those algorithms are usually intended for data that are sequentially generated by a mechanism whose behaviour changes over time. In sharp contrast, our switch distribution is especially suitable for situations where data are sampled from a *fixed* (though perhaps non-i.i.d.) source after all; the fact that one model temporarily leads to better predictions than another is caused by the fact that each "expert" $\bar{p}_k$ has itself already been designed as a universal predictor/estimator relative to some large set of distributions $\mathcal{M}_k$. The elements of $\mathcal{M}_k$ may be viewed as "base" predictors/experts, and the $\bar{p}_k$ may be thought of as meta-experts/predictors. Because of this two-stage structure, which meta-predictor $\bar{p}_k$ is best changes over time, even though the optimal base-predictor $\arg\min_{p\in\mathcal{M}} R_n(p^*, p)$ does not change over time.

If one of the considered prediction strategies $\bar{p}_k$ makes the best predictions eventually, our goal is to achieve consistent model selection: the total number of switches should also remain bounded. To this end we have defined the switch distribution such that positive prior probability is associated with switching finitely often and thereafter using $\bar{p}_k$ for all further outcomes. We need this property to prove that our method is consistent. Other dynamic expert tracking algorithms, such as the FixedShare algorithm [46], have been designed with different goals in mind, and as such they do not have this property. Not surprisingly then, our results do not resemble any of the existing results in the "tracking"-literature.

5.7 The Catch-Up Phenomenon, Bayes and Cross-Validation

5.7.1 The Catch-Up Phenomenon is Unbelievable! (according to p_{bma})

On page 122 we introduced the marginal Bayesian distribution $p_{\mathrm{bma}}(x^n) := \sum_k w(k)\bar{p}_k(x^n)$. If the distributions $\bar{p}_k$ are themselves Bayesian marginal distributions as in (5.1), then p_{bma} may be interpreted as (the density corresponding to) a distribution on the data that reflects some prior beliefs about the domain that is being modelled, as represented by the priors $w(k)$ and $w_k(\theta)$. If $w(k)$ and $w_k(\theta)$ truly reflected some decision-maker's a priori beliefs, then it is clear that the decision-maker would like to make sequential predictions of X_{n+1} given $X^n = x^n$ based on p_{bma} rather than on p_{sw}. Indeed, as we now show, the catch-up phenomenon as depicted in Figure 5.1 is exceedingly unlikely to take place under p_{bma}, and *a priori* a subjective Bayesian should be prepared to bet a lot of money that it does not

occur. To see this, consider the *no-hypercompression inequality* [39], versions of which are also known as "Barron's inequality" [5] and "competitive optimality of the Shannon-Fano code" [25]. It states that for any two distributions P and Q for X^∞, the P-probability that Q outperforms P by k bits or more when sequentially predicting $X_1, X_2, \ldots$ is exponentially small in k: for each n,

$$P(-\log q(X^n) \leq -\log p(X^n) - k) \leq 2^{-k}.$$

Plugging in $p_{\rm bma}$ for p, and $p_{\rm sw}$ for q, we see that what happened in Figure 5.1 ($p_{\rm sw}$ outperforming $p_{\rm bma}$ by about 40000 bits) is an event with probability no more than 2^{-40000} according to $p_{\rm bma}$. Yet, in many practical situations, the catch-up phenomenon does occur and $p_{\rm sw}$ gains significantly compared to $p_{\rm bma}$. This can only be possible because either the models are wrong (clearly, The Picture of Dorian Gray has not been drawn randomly from a finite-order Markov chain), or because the priors are "wrong" in the sense that they somehow don't match the situation one is trying to model. For this reason, some subjective Bayesians, when we confronted them with the catch-up phenomenon, have argued that it is just a case of "garbage in, garbage out" (GIGO): when the phenomenon occurs, then, rather than using the switch-distribution, one should reconsider the model(s) and prior(s) one wants to use, and, once one has found a superior model $\mathcal{M}'$ and prior w', one should use $p_{\rm bma}$ relative to $\mathcal{M}'$ and w'. Of course we agree that *if* one can come up with better models, one should of course use them. Nevertheless, we strongly disagree with the GIGO point of view: We are convinced that in practice, "correct" priors may be impossible to obtain; similarly, people are forced to work with "wrong" models all the time. In such cases, rather than embarking on a potentially never-ending quest for better models, the hurried practitioner may often prefer to use the imperfect – yet still useful – models that he has available, *in the best possible manner*. And then it makes sense to use $p_{\rm sw}$ rather than the Bayesian $p_{\rm bma}$: the best one can hope for in general is to regard the distributions in one's models as prediction strategies, and try to predict as well as the best strategy contained in any of the models, and $p_{\rm sw}$ is better at this than $p_{\rm bma}$. Indeed, the catch-up phenomenon raises some interesting questions for Bayes factor model selection: no matter what the prior is, by the no-hypercompression inequality above with $p = p_{\rm bma}$ and $q = p_{\rm sw}$, when comparing two models $\mathcal{M}_1$ and $\mathcal{M}_2$, before seeing any data, a Bayesian *always* believes that the switch-distribution will not substantially outperform $p_{\rm bma}$, which implies that a Bayesian *cannot* believe that, with non-negligible probability, a complex model $\bar{p}_2$ can at first predict substantially worse than a simple model $\bar{p}_1$ and then, for large samples, can predict substantially better. Yet in practice, this happens all the time!

5.7.2 Nonparametric Bayes

A more interesting subjective Bayesian argument against the switch distribution would be that, in the nonparametric setting, the data are sampled from some

$P^* \in \mathcal{M}^* \setminus \mathcal{M}$, and is not contained in any of the parametric models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ Yet, 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, our prior distribution is not really suitable for the situation that we are trying to model in the nonparametric setting, and we should use a nonparametric prior instead. While we completely agree with this reasoning, we would immediately like to add that the question then becomes: what nonparametric prior *should* one use? Nonparametric Bayes has become very popular in recent years, and it often works surprisingly well. Still, its practical and theoretical performance strongly depends on the type of priors that are used, and it is often far from clear what prior to use in what situation. In some situations, some nonparametric priors achieve optimal rates of convergence, but others can even make Bayes inconsistent [31, 39]. The advantage of the switch-distribution is that it does not require any difficult modelling decisions, but nevertheless under reasonable conditions it achieves the optimal rate of convergence in nonparametric settings, and, in the special case where one of the models on the list in fact approximates the true source extremely well, this model will in fact be identified (Theorem 5.2.1). In fact, one may think of p_{sw} as specifying a very special kind of nonparametric prior, and under this interpretation, our results are in complete agreement with the nonparametric Bayesian view.

5.7.3 Leave-One-Out Cross-Validation

From the other side of the spectrum, it has sometimes been argued that consistency is irrelevant, since in practical situations, the true distribution is never in any of the models under consideration. Thus, it is argued, one should use AIC-type methods such as leave-one-out cross-validation, because of their predictive optimality. We strongly disagree with this argument, for several reasons: first, in practical model selection problems, one is often interested in questions such as "does Y depend on feature X_k or not?" For example, $\mathcal{M}_{k-1}$ is a set of conditional distributions in which Y is independent of X_k, and $\mathcal{M}_k$ is a superset thereof in which Y can be dependent on X_k. There are certainly real-life situations where some variable X_j is truly completely irrelevant for predicting Y, and it may be the primary goal of the scientist to find out whether or not this is the case. In such cases, we would hope our model selection criterion to select, for large n, $\mathcal{M}_{k-1}$ rather than $\mathcal{M}_k$, and the problem with the AIC-type methods is that, because of their inconsistency, they sometimes do not do this. In other words, we think that consistency does matter, and we regard it as a clear advantage of the switch-distribution that it is consistent.

A second advantage over leave-one-out cross-validation is that the switch-distribution, like Bayesian methods, satisfies Dawid's *weak prequential principle* [29, 39]: the switch-distribution assesses the quality of a predictor $\bar{p}_k$ only in

terms of the quality of predictions *that were actually made*. To apply LOO on a sample $x_1, \ldots, x_n$, one needs to know the prediction for x_i given $x_1, \ldots, x_{i-1}$, but also $x_{i+1}, \ldots, x_n$. In practice, these may be hard to compute, unknown or even unknowable. An example of the first are non-i.i.d. settings such as time series models. An example of the second is the case where the $\bar{p}_k$ represent, for example, weather forecasters, or other predictors which have been designed to predict the future given the past. Actual weather forecasters use computer programs to predict the probability that it will rain the next day, given a plethora of data about air pressure, humidity, temperature etc. and the pattern of rain in the past days. It may simply be impossible to apply those programs in a way that they predict the probability of rain today, given data about tomorrow.

5.8 Conclusion and Future Work

We have identified the catch-up phenomenon as the underlying reason for the slow convergence of Bayesian model selection and averaging. Based on this, we have defined the switch-distribution P_{sw}, a modification of the Bayesian marginal distribution which is consistent, but also under some conditions achieves a minimax optimal convergence rate, a significant step forward in resolving the the AIC/BIC dilemma. Different strands of future work suggest themselves:

1. Lemma 5.3.5 provides a tool to prove minimax optimal in-sum convergence of the switch-distribution for particular nonparametric model classes $\mathcal{M}^*$. However, because of time constraints we have currently only applied this to histogram density estimation. We hope to eventually show that the switch-distribution actually achieves the minimax optimal convergence rate for a wide class of nonparametric problems.

2. Since p_{sw} can be computed in practice, the approach can readily be tested with real and simulated data in both density estimation and regression problems. Initial results on simulated data, on which we will report elsewhere, give empirical evidence that p_{sw} behaves remarkably well in practice. Model selection based on p_{sw}, like for p_{bma}, typically identifies the true distribution at moderate sample sizes. Prediction and estimation based on P_{sw} is of comparable quality to leave-one-out cross-validation (LOO) and generally, in no experiment did we find that it behaved substantially worse than either LOO or AIC.

3. It is an interesting open question whether analogues of Lemma 5.3.5 and Theorem 5.3.7 exist for model *selection* rather than averaging. In other words, in settings such as histogram density estimation where model averaging based on the switch distribution achieves the minimax convergence rate, does model selection based on the switch distribution achieve it as

well? For example, in Figure 5.1, sequentially predicting by the $\bar{p}_{K_{n+1}}$ that has maximum a posteriori probability (MAP) under the switch-distribution given data x^n, 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 40 000 bits. In the experiments mentioned above, we invariably found that predicting by the MAP $\bar{p}_{K_{n+1}}$ empirically converges at the same rate as using model averaging, i.e. predicting by $P_{\rm sw}$. However, we have no proof that this really must always be the case. Analogous results in the MDL literature suggest that a theorem bounding the risk of switch-based model selection, if it can be proven at all, would bound the squared Hellinger rather than the KL risk ([39], Chapter 15).

4. The way we defined $P_{\rm sw}$, it does not seem suitable for situations in which the number of considered models or model combinations is exponential in the sample size. Because of condition (i) in Lemma 5.3.5, our theoretical results do not cover this case either. Yet this case is highly important in practice, for example, in the subset selection problem [101]. It seems clear that the catch-up phenomenon can and will also occur in model selection problems of that type. Can our methods be adapted to this situation, while still keeping the computational complexity manageable? And what is the relation with the popular and computationally efficient L_1-approaches to model selection? [90]

5.9 Proofs

5.9.1 Proof of Theorem 5.2.1

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

$$\lim_n \frac{\sum_{\mathbf{s}\in U_n} \pi\big(\mathbf{s}\big) q_{\mathbf{s}}(X^n)}{\sum_{\mathbf{s}\in\mathbb{S}} \pi\big(\mathbf{s}\big) q_{\mathbf{s}}(X^n)} = 0 \qquad \text{with } \bar{P}_{k^*}\text{-probability 1.} \tag{5.37}$$

To see this, suppose the theorem is false. Then there exists a $\Phi \subseteq \Theta_{k^*}$ with $w_{k^*}(\Phi) := \int_\Phi w_{k^*}(\theta) d\theta > 0$ such that (5.6) does not hold for any $\theta^* \in \Phi$. But then by definition of $\bar{P}_{k^*}$ we have a contradiction with (5.37).

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

$$\sum_{\mathbf{s}'\in U_n} \pi(\mathbf{s}') q_{\mathbf{s}'}(x^n) \;\leq\; \sum_{\mathbf{s}\in A} \big(\pi(\mathbf{s}) + \pi(E_{\mathbf{s}})\big)\, q_{\mathbf{s}}(x^n) \;\leq\; (1+c) \sum_{\mathbf{s}\in A} \pi(\mathbf{s}) q_{\mathbf{s}}(x^n). \tag{5.38}$$

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

$$\lim_n \frac{r(X^n)}{\pi(\mathbf{s} = (0, k^*)) \cdot \bar{p}_{k^*}(X^n)} = 0 \qquad \text{with } \bar{P}_{k^*}\text{-probability 1.} \tag{5.39}$$

Using (5.38) 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 (5.37).

For all $\mathbf{s} \in A$ and $x^{t_m(\mathbf{s})} \in \mathcal{X}^{t_m(\mathbf{s})}$, by definition $Q_{\mathbf{s}}(X_{t_m+1}^{\infty}|x^{t_m})$ is equal to $\bar{P}_{k_m}(X_{t_m+1}^{\infty}|x^{t_m})$, 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. In that case, the proof follows by Corollary 5.9.2 proved in Section 5.9.3. Any countable mixture of distributions that are mutually singular with P_{k^*}, in particular R, is mutually singular with P_{k^*}. This implies (5.39) by Lemma 3.1 of [5], which says that for any two mutually singular distributions R and P, the density ratio $r(X^n)/p(X^n)$ goes to zero as $n \to \infty$ with P-probability 1. □

5.9.2 Proof of Theorem 5.2.2

The proof is almost identical to the proof of Theorem 5.2.1. 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

$$\lim_n \frac{\sum_{\mathbf{s}\in U_n} \pi\big(\mathbf{s}\big) q_{\mathbf{s}}(X^n)}{\sum_{\mathbf{s}\in\mathbb{S}} \pi\big(\mathbf{s}\big) q_{\mathbf{s}}(X^n)} = 0 \qquad \text{with } \bar{P}_{k^*}^{\mathrm{B}}\text{-probability 1.} \tag{5.40}$$

Note that the $q_{\mathbf{s}}$ in (5.40) are defined relative to the non-Bayesian estimators $\bar{p}_1, \bar{p}_2, \ldots$, whereas the $\bar{P}_{k^*}^{\mathrm{B}}$ on the right of the equation is the probability according to a *Bayesian* marginal distribution $\bar{P}_{k^*}^{\mathrm{B}}$, which has been chosen so that the theorem's condition holds. To see that (5.40) is sufficient to prove the theorem, suppose the theorem is false. Then, because the prior w_{k^*} is mutually absolutely continuous with Lebesgue measure, there exists a $\Phi \subseteq \Theta_{k^*}$ with nonzero prior measure under w_{k^*}, such that (5.8) does not hold for any $\theta^* \in \Phi$. But then by definition of $\bar{P}_{k^*}^{\mathrm{B}}$ we have a contradiction with (5.40).

Using exactly the same reasoning as in the proof of Theorem 5.2.1, it follows that, analogously to (5.39), we have

$$\lim_n \frac{r(X^n)}{\pi(\mathbf{s} = (0, k^*)) \cdot \bar{p}_{k^*}^{\mathrm{B}}(X^n)} = 0 \qquad \text{with } \bar{P}_{k^*}^{\mathrm{B}}\text{-probability 1.} \tag{5.41}$$

This is just (5.39) with r now referring to a mixture of switch distributions defined relative to the non-Bayesian estimators $\bar{p}_1, \bar{p}_2, \ldots$, and the $\bar{p}_{k^*}^{\mathrm{B}}$ in the denominator

and on the right referring to the Bayesian marginal distribution $\bar{P}^{\mathrm{B}}_{k^*}$. Using (5.38) and the fact that $\sum_{\mathbf{s}\in\mathbb{S}} \pi(\mathbf{s}) q_{\mathbf{s}}(x^n) \geq \pi(\mathbf{s} = (0, k^*)) \cdot \bar{p}_{k^*}(x^n)$, and the fact that, by assumption, for some K, for all large n, $\bar{p}_{k^*}(X^n) \geq \bar{p}^{\mathrm{B}}_{k^*}(X^n) 2^{-K}$ with $\bar{P}^{\mathrm{B}}_{k^*}$-probability 1, (5.41) implies (5.40). □

5.9.3 Mutual Singularity as Used in the Proof of Theorem 5.2.1

Let $Y^2 = (Y_1, Y_2)$ be random variables that take values in separable metric spaces Ω_1 and Ω_2, respectively. We will assume all spaces to be equipped with Borel σ-algebras generated by the open sets. Let p be a prediction strategy for Y^2 with corresponding distributions $P(Y_1)$ and, for any $y^1 \in \Omega_1$, $P(Y_2|y^1)$. To ensure that $P(Y^2)$ is well-defined, we impose the requirement that for any fixed measurable event $A_2 \subseteq \Omega_2$ the probability $P(A_2|y^1)$ is a measurable function of y^1.

Lemma 5.9.1. *Suppose p and q are prediction strategies for $Y^2 = (Y_1, Y_2)$, which take values in separable metric spaces Ω_1 and Ω_2, respectively. Then if $P(Y_2|y^1)$ and $Q(Y_2|y^1)$ are mutually singular for all $y^1 \in \Omega_1$, then $P(Y^2)$ and $Q(Y^2)$ are mutually singular.*

The proof, due to Peter Harremoës, is given below the following corollary, which is what we are really interested in. Let $X^\infty = X_1, X_2, \ldots$ be random variables that take values in the separable metric space $\mathcal{X}$. Then what we need in the proof of Theorem 5.2.1 is the following corollary of Lemma 5.9.1:

Corollary 5.9.2. *Suppose p and q are prediction strategies for the sequence of random variables $X^\infty = X_1, X_2, \ldots$ that take values in respective separable metric spaces $\mathcal{X}_1, \mathcal{X}_2, \ldots$ Let m be any positive integer. Then if $P(X^\infty_{m+1}|x^m)$ and $Q(X^\infty_{m+1}|x^m)$ are mutually singular for all $x^m \in \mathcal{X}^m$, then $P(X^\infty)$ and $Q(X^\infty)$ are mutually singular.*

Proof. The product spaces $\mathcal{X}_1 \times \cdots \times \mathcal{X}_m$ and $\mathcal{X}_{m+1} \times \mathcal{X}_{m+2} \times \cdots$ are separable metric spaces [64, pp. 5,6]. Now apply Lemma 5.9.1 with $\Omega_1 = \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$ and $\Omega_2 = \mathcal{X}_{m+1} \times \mathcal{X}_{m+2} \times \cdots$. □

Proof of Lemma 5.9.1. For each $\omega_1 \in \Omega_1$, by mutual singularity of $P(Y_2|\omega_1)$ and $Q(Y_2|\omega_1)$ there exists a measurable set $C_{\omega_1} \subseteq \Omega_2$ such that $P(C_{\omega_1}|\omega_1) = 1$ and $Q(C_{\omega_1}|\omega_1) = 0$. As Ω_2 is a metric space, it follows from [64, Theorems 1.1 and 1.2 in Chapter II] that for any $\epsilon > 0$ there exists an open set $U^\epsilon_{\omega_1} \supseteq C_{\omega_1}$ such that

$$P(U^\epsilon_{\omega_1}|\omega_1) = 1 \quad \text{and} \quad Q(U^\epsilon_{\omega_1}|\omega_1) < \epsilon. \tag{5.42}$$

As Ω_2 is a separable metric space, there also exists a countable sequence $\{B_i\}_{i\geq 1}$ of open sets such that every open subset of Ω_2 ($U^\epsilon_{\omega_1}$ in particular) can be expressed as the union of sets from $\{B_i\}$ [64, Theorem 1.8 in Chapter I].

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

$$P(V^\epsilon_{\omega_1}|\omega_1) > 1-\epsilon \quad \text{and} \quad Q(V^\epsilon_{\omega_1}|\omega_1) < \epsilon. \tag{5.43}$$

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

$$A^\epsilon_i = \{\omega_1 \in \Omega_1 : P(D_i|\omega_1) > 1-\epsilon, Q(D_i|\omega_1) < \epsilon\} \setminus \bigcup_{j=1}^{i-1} A^\epsilon_j \tag{5.44}$$

for $i = 1, 2, \ldots$ Note that, by the reasoning above, for each $\omega_1 \in \Omega_1$ there exists an i such that $\omega_1 \in A^\epsilon_i$, which implies that $\{A^\epsilon_i\}$ forms a partition of Ω_1. Now, as all elements of $\{A^\epsilon_i\}$ and $\{D_i\}$ are measurable, so is the set $F^\epsilon = \bigcup_{i=1}^\infty A^\epsilon_i \times D_i \subseteq \Omega_1 \times \Omega_2$, for which we have that $P(F^\epsilon) = \sum_{i=1}^\infty P(A^\epsilon_i \times D_i) > (1-\epsilon)\sum_{i=1}^\infty P(A_i) = 1-\epsilon$ and likewise $Q(F^\epsilon) < \epsilon$.

Finally, let $G = \bigcap_{n=1}^\infty \bigcup_{k=n}^\infty F^{2^{-k}}$. Then $P(G) = \lim_{n\to\infty} P(\bigcup_{k=n}^\infty F^{2^{-k}}) \geq \lim_{n\to\infty} 1 - 2^{-n} = 1$ and $Q(G) = \lim_{n\to\infty} Q(\bigcup_{k=n} F^{2^{-k}}) \leq \lim_{n\to\infty} \sum_{k=n}^\infty 2^{-k} = \lim_{n\to\infty} 2^{-n+1} = 0$, which proves the lemma. □

5.9.4 Proof of Theorem 5.5.1

Before we prove Theorem 5.5.1, we first need to establish some additional properties of the prior π as defined in (5.7). Define, for all $n \in \mathbb{N}$ and $\mathbf{s} = ((t_1, k_1), \ldots, (t_m, k_m)) \in \mathbb{S}$:

$$S_n(\mathbf{s}) := \mathbf{1}_{\{t_1,\ldots,t_m\}}(n-1); \tag{5.45}$$
$$M_n(\mathbf{s}) := \mathbf{1}_{\{t_m, t_m+1,\ldots\}}(n-1); \tag{5.46}$$
$$K_n(\mathbf{s}) := k_i \text{ for the unique } i \text{ such that } t_i < n \text{ and } i = m \vee t_{i+1} \geq n. \tag{5.47}$$

These functions denote, respectively, whether or not a switch occurs just before outcome n, whether or not the last switch occurs somewhere before outcome n and which prediction strategy is used for outcome n. The prior π determines the distributions of these random variables. We also define $E_n(\mathbf{s}) := (S_n(\mathbf{s}), M_n(\mathbf{s}), K_n(\mathbf{s}))$ as a convenient abbreviation. Every parameter value $\mathbf{s} \in \mathbb{S}$ induces an infinite sequence of values $E_1, E_2, \ldots$ The advantage of these new variables is that they allow us to reformulate the prior as a strategy for prediction of the value of the next random variable E_{n+1} (which in turn determines the distribution on X_{n+1} given x^n), given all previous random variables E^n. Therefore, we first calculate

the conditional probability $\pi(E_{n+1}|E^n)$ before proceeding to prove the theorem. As it turns out, our prior has the nice property that this conditional probability has a very simple functional form: depending on E_n, it is either zero, or a function of only E_{n+1} itself. This will greatly facilitate the analysis.

Lemma 5.9.3. *Let* $\pi(\mathbf{s}) = \theta^{m-1}(1-\theta)\pi_{\text{K}}(k_1)\prod_{i=2}^{m}\pi_{\text{T}}(t_i|t_i > t_{i-1})\pi_{\text{K}}(k)$ *as in* (5.7). *For* $\pi(\mathbf{s}) > 0$ *we have*

$$\pi(E_1) := \pi_{\text{K}}(K_1)\begin{cases}\theta & \text{if } M_1 = 0\\ 1-\theta & \text{if } M_1 = 1\end{cases} \tag{5.48}$$

$$\pi(E_{n+1}|E^n) := \begin{cases}\pi_{\text{T}}(Z > n|Z \geq n) & \text{if } S_{n+1} = 0 \text{ and } M_{n+1} = 0\\ 1 & \text{if } S_{n+1} = 0 \text{ and } M_{n+1} = 1\\ \pi_{\text{T}}(Z = n|Z \geq n)\pi_{\text{K}}(K_{n+1})\theta & \text{if } S_{n+1} = 1 \text{ and } M_{n+1} = 0\\ \pi_{\text{T}}(Z = n|Z \geq n)\pi_{\text{K}}(K_{n+1})(1-\theta) & \text{if } S_{n+1} = 1 \text{ and } M_{n+1} = 1.\end{cases} \tag{5.49}$$

Proof. To check (5.48), note that we must have either $E_1 = (1, 1, k)$, which corresponds to $\mathbf{s} = (0, k)$ which has probability $\pi_{\text{K}}(k)(1-\theta)$ as required, or $E_1 = (1, 0, k)$. The latter corresponds to the event that $m > 1$ and $k_1 = k$, which has probability $\pi_{\text{K}}(k)\theta$.

We proceed to calculate the conditional distribution $\pi(E_{n+1}|E^n)$. We distinguish the case that $M_n(\mathbf{s}) = 1$ (the last switch defined by $\mathbf{s}$ occurs before sample size n), and $M_n(\mathbf{s}) = 0$ (there will be more switches). First suppose $M_n(\mathbf{s}) = 0$, and let $a_n = \max\{i \mid t_i < n\} = \sum_{i=1}^{n} S_i$. Then

$$\begin{aligned}\pi(E^n) &= \sum_{m=a_n+1}^{\infty}\sum_{t_{a_n+1}=n}\sum_{\substack{t_{a_n+2},\ldots,t_m\\ k_{a_n+1},\ldots,k_m}}\pi_{\text{M}}(m)\prod_{i=1}^{m}\pi_{\text{T}}(t_i|t_i > t_{i-1})\pi_{\text{K}}(k_i)\\ &= \sum_{m=a_n+1}^{\infty}\pi_{\text{M}}(m)\left(\prod_{i=1}^{a_n}\pi_{\text{T}}(t_i|t_i > t_{i-1})\pi_{\text{K}}(k_i)\right)\sum_{t_{a_n+1}=n}^{\infty}\pi_{\text{T}}(t_{a_n+1}|t_{a_n+1} > t_{a_n})\\ &\cdot\sum_{t_{a_n+2},\ldots,t_m}\left(\prod_{i=a_n+2}^{m}\pi_{\text{T}}(t_i|t_i > t_{i-1})\right)\sum_{k_{a_n+1},\ldots,k_m}\prod_{i=a_n+1}^{m}\pi_{\text{K}}(k_i)\\ &= \pi_{\text{M}}(Z > a_n)\left(\prod_{i=1}^{a_n}\pi_{\text{T}}(t_i|t_i > t_{i-1})\pi_{\text{K}}(k_i)\right)\pi_{\text{T}}(t_{a_n+1} \geq n)\cdot 1\cdot 1.\end{aligned} \tag{5.50}$$

If $M_n(\mathbf{s}) = 1$, then there is only one $\mathbf{s}$ that matches E^n, which has probability

$$\pi(E^n) = \pi_{\text{M}}(Z = a_n) \prod_{i=1}^{a_n} \pi_{\text{T}}(t_i | t_i > t_{i-1}) \pi_{\text{K}}(k_i). \tag{5.51}$$

From (5.50) and (5.51) we can compute the conditional probability $\pi(E_{n+1}|E^n)$. We distinguish further on the basis of the possible values of S_{n+1} and M_{n+1}, which together determine M_n (namely, if $M_{n+1} = 0$ then $M_n = 0$ and if $M_{n+1} = 1$ then $M_n = 1 - S_{n+1}$). Also note that $S_{n+1} = 0$ implies $a_{n+1} = a_n$ and $S_{n+1} = 1$ implies $a_{n+1} = a_n + 1$ and $t_{a_n+1} = n$. Conveniently, most factors cancel out, and we obtain

$$\begin{aligned}
&\pi(E_{n+1}|E^n) \\
&= \begin{cases}
\pi_{\text{T}}(t_{a_{n+1}} + 1 \geq n+1) / \pi_{\text{T}}(t_{a_n} \geq n) & \text{if } S_{n+1} = 0,\ M_{n+1} = 0 \\
1 & \text{if } S_{n+1} = 0,\ M_{n+1} = 1 \\
\frac{\pi_{\text{M}}(Z > a_n+1)}{\pi_{\text{M}}(Z > a_n)} \pi_{\text{T}}(t_{a_n+1} | t_{a_n+1} > t_{a_n}) \pi_{\text{K}}(k_{a_n+1}) \frac{\pi_{\text{T}}(t_{a_n+2} \geq n+1)}{\pi_{\text{T}}(t_{a_n+1} \geq n)} & \text{if } S_{n+1} = 1,\ M_{n+1} = 0 \\
\frac{\pi_{\text{M}}(Z = a_n+1)}{\pi_{\text{M}}(Z > a_n)} \frac{\pi_{\text{T}}(t_{a_n+1} | t_{a_n+1} > t_{a_n})}{\pi_{\text{T}}(t_{a_n+1} \geq n)} \pi_{\text{K}}(k_{a_n+1}) & \text{if } S_{n+1} = 1,\ M_{n+1} = 1,
\end{cases}
\end{aligned}$$

which reduces to (5.49). □

Proof of Theorem 5.5.1. We will use a number of independence properties of P in this proof. First, we have that the distribution on E_{n+1} is independent of X^n conditional on E^n, because, using Bayes' rule,

$$\begin{aligned}
P(E_{n+1}|E^n, X^n) &= \frac{P(X^n|E^{n+1}) P(E_{n+1}|E^n)}{P(X^n|E^n)} \\
&= \frac{P(X^n|E^n) P(E_{n+1}|E^n)}{P(X^n|E^n)} = \pi(E_{n+1}|E^n),
\end{aligned} \tag{5.52}$$

provided that $P(E^{n+1}, X^n) > 0$. In turn, whether or not E_{n+1} can occur depends only on M_n and K_n. For all $n \geq 1$, define the function $N(M_n, K_n)$ as the set of values of the E_{n+1} that have positive probability conditional on M_n and K_n, i.e.

$$N(M_n, K_n) := \{(s, m, k) \mid \text{either } s = 1 \wedge M_n = 0 \text{ or } s = 0 \wedge m = M_n \wedge k = K_n\}. \tag{5.53}$$

Thus, the conditional distribution on E_{n+1} given all previous values E^n and all observations X^n is a function of only E_{n+1} itself, n, M_n and K_n. It remains the same whether or not any of the other variables are included in the conditional. If $S_{n+1} = 1$ then it is not even a function of K_n. This interesting property is used three times in the following. Namely,

1. Since $(0, 0, k) \in N(0, k)$, we have $\pi_{\text{T}}(Z > n | Z \geq n) = P(E_{n+1} = (0, 0, k) \,|\, x^n, M_n = 0, K_n = k)$.

2. Since $(1,0,k) \in N(0,k')$ for all k', we have $\pi_{\text{T}}(Z = n|Z \geq n)\pi_{\text{K}}(k)\theta = P(E_{n+1} = (1,0,k) \,|\, x^n, M_n = 0)$.

3. If $k \in \mathcal{K}_1$, then $\pi_{\text{K}}(k)\theta = P(x^0, M_1 = 0, K_1 = k)$.

We first show that the invariants $w_k^a = P(x^{n-1}, M_n = 0, K_n = k)$ and $w_k^b = P(x^{n-1}, M_n = 1, K_n = k)$ hold at the start of each iteration (before line 3). The invariants ensure that $w_k^a + w_k^b = P(x^{n-1}, K_n = k)$ so that the correct probabilities are reported.

Line 1 initialises w_k^a to $\theta\pi_{\text{K}}(k)$ for $k \in \mathcal{K}_1$. By item 3 this equals $P(x^0, M_1 = 0, K_1 = k)$ as required. We omit calculations for w_k^b, which run along the same lines as for w_k^a. Thus the loop invariant holds at the start of the first iteration.

We proceed to go through the algorithm step by step to show that the invariant holds in subsequent iterations as well. In the loss update in line 4 we update the weights for $k \in \mathcal{K}_n$ to

$$\begin{aligned} w_k^a &= P(x^{n-1}, M_n = 0, K_n = k) \cdot p_k(x_n|x^{n-1}) \\ &= \sum_{\mathbf{s}: M_n = 0, K_n = k} \pi(\mathbf{s}) \left(\prod_{i=1}^{n-1} p_{K_i}(x_i|x^{i-1}) \right) p_{K_n}(x_n|x^{n-1}) \;=\; P(x^n, M_n = 0, K_n = k). \end{aligned}$$

Similarly $w_k^b = P(x^n, M_n = 1, K_n = k)$. Then in line 5, we compute $\texttt{pool} = \pi_{\text{T}}(Z = n|Z \geq n) \sum_{k \in \mathcal{K}_n} P(x^n, M_n = 0, K_n = k) = \pi_{\text{T}}(Z = n|Z \geq n) P(x^n, M_n = 0)$.

Finally, after the loop that starts at line 6 and ends at line 9, we obtain for all $k \in \mathcal{K}_{n+1}$:

$$\begin{aligned} w_k^a &= P(x^n, M_n = 0, K_n = k)\pi_{\text{T}}(Z > n|Z \geq n) + \pi_{\text{T}}(Z = n|Z \geq n) P(x^n, M_n = 0)\pi_{\text{K}}(k)\theta \\ &= P(x^n, M_n = 0, K_n = k) P(S_{n+1} = 0, M_{n+1} = 0 | x^n, M_n = 0, K_n = k) \\ &+ \; P(x^n, M_n = 0) P(S_{n+1} = 1, M_{n+1} = 0, K_{n+1} = k | x^n, M_n = 0) \\ &= P(x^n, M_n = 0, K_n = k, S_{n+1} = 0, M_{n+1} = 0) \\ &+ \; P(x^n, M_n = 0, S_{n+1} = 1, M_{n+1} = 0, K_{n+1} = k) \\ &= P(x^n, S_{n+1} = 0, M_{n+1} = 0, K_{n+1} = k) + P(x^n, S_{n+1} = 1, M_{n+1} = 0, K_{n+1} = k) \\ &= P(x^n, M_{n+1} = 0, K_{n+1} = k). \end{aligned}$$

Here we used items 1 and 2 in the second equality. Again, a similar derivation shows that $w_k^b = P(x^n, K_{n+1} = k, M_{n+1} = 1)$. These weights satisfy the invariant at the beginning of the next iteration; after the last iteration the final posterior is also correctly reported based on these weights. □