Universality of Bayesian Predictions

Alessio Sancetta*
University of Cambridge

July 23, 2007

Abstract

Given the sequential update nature of Bayes rule, Bayesian methods find natural application to prediction problems. Advances in computational methods allow to routinely use Bayesian methods in econometrics. Hence, there is a strong case for feasible predictions in a Bayesian framework. This paper studies the theoretical properties of Bayesian predictions and shows that under minimal conditions we can derive finite sample bounds for the loss incurred using Bayesian predictions under the Kullback-Leibler divergence. In particular, the concept of universality of predictions is discussed and universality is established for Bayesian predictions in a variety of settings. These include predictions under almost arbitrary loss functions, model averaging, predictions in a non stationary environment and under model miss-specification. Given the possibility of regime switches and multiple breaks in economic series, as well as the need to choose among different forecasting models, which may inevitably be miss-specified, the finite sample results derived here are of interest to economic and financial forecasting.

Key Words: Bayesian prediction, model averaging, universal prediction.

JEL Classification : C11, C44, C53.

1 Introduction

Bayesian methods have gained increasing importance in empirical work. In this respect, macro-policy modelling is one of its success story. Indeed highly dimensional macroeconometric models are often estimated an analyzed within a Bayesian framework (e.g. Sims and Zha, 1998, and the reviews of An and Schorfheide, 2007, and Schorfheide 2007, where many references can be found). Besides large dimensional macro-models used for policy making, there are many applications of Bayesian methods to econometrics problems with strong empirical motivations related to macroeconomic and financial forecasting (e.g. Canova and Ciccarelli, 2004, Pesaran et al., 2006, Chib et al., 2006).

[*] **Address for Correspondence:** Faculty of Economics, Sidgwick Avenue, University of Cambridge, Cambridge CB3 9DD, UK. E-mail: alessio.sancetta@econ.cam.ac.uk.

The goal of these application is to infer something about the future from past information, when interest goes beyond point prediction. Motivated by the prediction problem, we will study the theoretical properties of Bayesian predictions which satisfy an important property called universality. The goal is to present general results about universality of Bayesian predictions. Some results are new, while others are known, though not necessarily in the form presented here and not in the econometric literature. All these results fall within the same unifying approach and their generality should induce the reader to consider the Bayesian approach as an ideal forecasting method. We consider optimal prediction under arbitrary loss function and optimal model averaging. We also consider the case when the optimal model changes over time and we wish to track these changes as much as possible. In these cases, the straight Bayesian update will not lead to a satisfactory prediction and some additional randomization over the models or parameters is required. Finally, we show that if the "true model" does not belong to the class of parametric models considered, the Bayesian predictor performs as well as the best parametric model in the class under no additional assumptions. Establishing a similar result in the maximum likelihood context would require more stringent conditions (e.g. Strasser, 1981, and Gourieroux et al., 1984, for results related to this claim, Phillips and Ploberger, 1996, for asymptotic connections between Bayes and maximum likelihood methods).

Improvements in computational power and the presence of a rich number of computational methods have made possible to routinely use Bayesian methods in practice (e.g. Chib, 2004, Evans and Swartz, 1995, Geweke, 1989, 2005). Moreover, results concerning dimensionality reduction may further alleviate the computational burden (e.g. Cardigan and Raftery, 1994, for Bayesian model averaging). Computational issues will not be discussed here and the interested reader should consult the above references.

Bayesian prediction is based on the natural principle that new collected evidence should be used to update predictions in a forecasting problem. Bayes rule satisfies optimality properties in terms of information processing (e.g. Zellner, 1988, 2002, Clarke, 2007) and Bayesian estimation requires weaker conditions for consistency than other methods like maximum likelihood estimation (e.g. Strasser, 1981). Predictions based on Bayes rule lead to forecasts that perform uniformly well over the whole parameter space. Forecasts satisfying this property will be called universal. This only requires a mild condition on the prior, i.e. the prior needs to be information dense at the "true value" (e.g. Barron, 1988, 1998). It is a remarkable fact that this condition is not sufficient for consistency of posterior distributions (e.g. Diaconis and Freedman, 1986, Barron, 1998).

There is a rich statistical literature on consistency of Bayesian procedures (e.g. Barron, 1998, for a survey) to which the results of this paper are related. However, the present discussion will also bring together ideas and results from a rich literature in information theory (e.g. Merhav and Feder, 1998), artificial intelligence (e.g. Cesa-Bianchi and Lugosi, 2005, Hutter 2005), and game theory (e.g. see special issue in Games and Economic Behavior, Vol. 29, 1999). It is not possible to provide a review of the results in all these areas. However, each the theorems stated here will be followed by a discussion of related references.

The focus of the paper is theoretical. However, its conclusions have clear practical implication for the use of Bayesian prediction and provide guidelines for the choice of prior. The choice of prior is not crucial as long as it satisfies some general conditions. Under additional smoothness conditions on the likelihood w.r.t. the unknown parameter, the optimal choice of prior is known to be related to the information matrix (i.e. an exponential tilt of Jeffries' prior) and more details can be given (Clarke and Barron, 1990, for exact conditions), but will not be discussed here.

While conducting inference to distinguish between two hypotheses, the posterior odd ratio represents the evidence in favor of one hypothesis relative to another. The posterior odd ratio is affected by the prior distribution. Hence, the Bayesian prediction and estimation problem contrasts with the testing problem, where the choice of prior is more crucial (e.g. Kass and Raftery, 1995, Section 5).

The plan of the paper is as follows. At first we provide background notation and definitions. We introduce the definition of universality of predictions and give a game theoretic justification for it, linking it to the prequential and real time econometrics literature. Section 2 states the universality results for a variety of problems including prediction under almost arbitrary loss function, model averaging, predictions in a non-stationary environment and predictions under miss-specification. Further discussion including remarks about the conditions can be found in Section 3. Proofs are in the appendix.

1.1 Background and Notation

For $t \in \mathbb{N}$, let $Z_1, ..., Z_t$ be random variables each taking values in some set $\mathcal{Z}$ and with joint law P_θ where $\theta \in \Theta$, for some set Θ. For ease of notation, we suppress the dependence of P_θ on t, the number of random variables. In particular $P_\theta\left(\bullet|\mathcal{F}_{t-1}\right)$ denotes the law of Z_t conditional on $\mathcal{F}_{t-1}$, where $\mathcal{F}_{t-1}$ is the sigma algebra generated by $\left(Z_s\right)_{s<t}$ and $\mathcal{F}_0$ is assumed to be trivial. It follows that

$$P_\theta\left(z_1^t\right) = \prod_{s=1}^{t} P_\theta\left(z_s|\mathcal{F}_{s-1}\right)$$

where $z_1^t := (z_1, ..., z_t)$ (where the above are understood as distribution functions). We assume that P_θ is absolutely continuous with respect to a sigma finite measure μ and define its density (w.r.t. μ) by p_θ. When $\theta \in \Theta$ is unknown, the Bayesian estimator of $p_\theta\left(z_1^t\right)$ is given by

$$p_w\left(z_1^t\right) = \int_\Theta p_\theta\left(z_1^t\right) w\left(d\theta\right)$$

where w is a prior probability measure on subsets of Θ. Note that if we assume Θ compact, then $\int_\Theta dw < \infty$ for any sigma finite measure w. Hence, if w is a diffuse prior on a Euclidean set Θ, then we shall assume Θ compact, so that we may always turn a sigma finite measure w into a probability measure by standardization.

Example 1 *Suppose w is a uniform prior on $\Theta \subset \mathbb{R}$, then we just have $w\left(d\theta\right) = d\theta/\left|\Theta\right|$, where*

$|\Theta| < \infty$ *is the Lebesgue measure of* Θ.

An estimator for $p_\theta\left(z_t|\mathcal{F}_{t-1}\right) = p_\theta\left(z_1^t\right)/p_\theta\left(z_1^{t-1}\right)$ is just

$$p_w\left(z_t|\mathcal{F}_{t-1}\right) = \frac{p_w\left(z_1^t\right)}{p_w\left(z_1^{t-1}\right)} \tag{1}$$

where $0/0 := 0$.

We are interested in sequential prediction of $p_\theta\left(z_t|\mathcal{F}_{t-1}\right)$ for $t = 1, 2, 3, ...$ which is recursively estimated as

$$p_w\left(z_t|\mathcal{F}_{t-1}\right) = \int_\Theta p_\theta\left(z_t|\mathcal{F}_{t-1}\right) w\left(d\theta|\mathcal{F}_{t-1}\right) \tag{2}$$

where

$$w\left(d\theta|\mathcal{F}_t\right) = \frac{w\left(d\theta|\mathcal{F}_{t-1}\right) p_\theta\left(Z_t|\mathcal{F}_{t-1}\right)}{\int_\Theta w\left(d\theta|\mathcal{F}_{t-1}\right) p_\theta\left(Z_t|\mathcal{F}_{t-1}\right)} \tag{3}$$

and $w\left(d\theta|\mathcal{F}_t\right)$ is the posterior probability written in sequential form, more commonly written as

$$w\left(d\theta|\mathcal{F}_t\right) = \frac{w\left(d\theta\right) p_\theta\left(Z_1^t\right)}{\int_\Theta w\left(d\theta\right) p_\theta\left(Z_1^t\right)}$$

where the above relations follow by induction. The justification of this approach is Bayes rule. In a prediction context, we shall quantify the sequential loss incurred by using $p_w\left(z_t|\mathcal{F}_{t-1}\right)$ instead of $p_\theta\left(z_t|\mathcal{F}_{t-1}\right)$. To this end, we shall use the Kullback-Leibler (KL) divergence

$$\begin{aligned} D_t\left(P_\theta\|P_w\right) &:= \int_{\mathcal{Z}} p_\theta\left(z|\mathcal{F}_{t-1}\right) \ln\left(\frac{p_\theta\left(z|\mathcal{F}_{t-1}\right)}{p_w\left(z|\mathcal{F}_{t-1}\right)}\right) \mu\left(dz\right) \\ &= \mathbb{E}_{t-1}^\theta\left[\ln\left(p_\theta\left(Z_t|\mathcal{F}_{t-1}\right)\right) - \ln\left(p_w\left(Z_t|\mathcal{F}_{t-1}\right)\right)\right] \end{aligned}$$

where $\mathbb{E}_t^\theta$ is expectation w.r.t. $P_\theta\left(\bullet|\mathcal{F}_{t-1}\right)$ and define $D_{1,T}\left(P_\theta\|P_w\right) := \sum_{t=1}^T D_t\left(P_\theta\|P_w\right)$ as the total KL divergence. KL divergence will be used interchangeably with the term relative entropy. We shall use $\mathbb{E}^\theta$ to denote unconditional expectation w.r.t. P_θ. Our interest is in predictions that are universal, as defined next.

Definition 1 *The prediction* p_w *is universal with respect to* $\{P_\theta : \theta \in \Theta\}$ *if*

$$\sup_{\theta\in\Theta} \frac{\mathbb{E}^\theta D_{1,T}\left(P_\theta\|P_w\right)}{T} \to 0$$

We now turn to the implications of universality.

1.2 Implications of Universality

Definition 1 has practical implications in a variety of contexts. For any prior w on Θ and any measure Q on $\mathcal{Z}^T$, the mutual information between w and Q is defined by

$$I(w, Q) := \int_\Theta \mathbb{E}^\theta D_{1,T}(P_\theta \| Q) w(d\theta)$$

(e.g. Clarke, 2007, Haussler and Opper, 1997). By the properties of the KL divergence, the mutual information is minimized w.r.t. Q by P_w, i.e.

$$I(w, P_w) \leq I(w, Q)$$

for any Q. Hence, the minimizer of the mutual information is the Bayes risk (e.g. Haussler and Opper, 1997, p. 2455). Universality of Bayesian prediction implies that the Bayes risk divided by T converges to zero.

The Bayes risk can be given a game theoretic interpretation. Suppose that the environment samples a $\theta \in \Theta$ according to the prior w and then observations Z_1^T are drawn according to P_θ. The forecaster only knows $\{P_{\theta'} : \theta' \in \Theta\}$ and that the prior is w. Then, a predictive distribution Q needs to be chosen such that the average loss $I(w, Q)$ is minimized.

Using universality, we can go a step further and consider the following adversarial game. Nature chooses $\theta \in \Theta$ such that $\mathbb{E}^\theta D_{1,T}(P_\theta \| Q)$ is maximized. The goal of the forecaster is to choose a predictive distribution Q such that $\sup_{\theta\in\Theta} \mathbb{E}^\theta D_{1,T}(P_\theta \| Q)$ is minimized. The solution to this problem is the Bayesian predictor P_w (Haussler, 1997, Theorem 1). Hence, the Bayesian prediction P_w solves the following minimax problem

$$\inf_Q \sup_{\theta\in\Theta} \mathbb{E}^\theta D_{1,T}(P_\theta \| Q)$$

where the inf is taken over all joint distributions Q on $\mathcal{Z}^T$.

Another important consequence of universality is in the context of prequential (predictive sequential) evaluation (e.g. Dawid, 1984, 1986). Dawid calls $D_{1,T}(P_\theta \| P_w)$ the prequential log-likelihood ratio. Given that $D_{1,T}(P_\theta \| P_w) \geq 0$, universality implies $L_1(P_\theta)$ convergence of the standardized prequential log-likelihood ratio, which in turn implies its convergence in P_θ-probability for any $\theta \in \Theta$. The prequential approach to statistical evaluation has also impact on real time econometric issues (Pesaran and Timmermann, 2005). It would be desirable to establish a.s. convergence of the prequential log-likelihood ratio. This is what the prequential approach advocates. Unfortunately, the method of proof used in this paper will not allow to do so. Note that expectation of the total relative entropy is equal to the relative entropy of the joint distributions.

The next question to ask is under what conditions on the prior universality holds. The sufficient condition for this is called information denseness and is discussed next.

1.3 Information Denseness and Resolvability Index

For any $\theta \in \Theta$, $T \in \mathbb{N}$, and $\delta > 0$, define the following set

$$B_T(\theta, \delta) := \left\{\theta' \in \Theta : \mathbb{E}^{\theta} D_{1,T}(P_{\theta} \| P_{\theta'}) \leq \delta\right\}. \tag{4}$$

To ease notation, we may write $B_T(\theta, \delta) = B_T(\theta)$ whichever is felt more appropriate for the situation. The set $B_T(\theta, \delta)$ is called information neighbor and is the set of subsets of Θ with expected total relative entropy less or equal to $\delta > 0$. Then, the prior w is said to be information dense (at θ) if it assign strictly positive probability to each information neighbor of size $\delta_T T$, i.e. $w(B_T(\theta, \delta_T T)) > 0$ for any $\delta_T > 0$. Information denseness of the prior is often used in the Bayesian consistency literature (e.g. Barron, 1998, Barron et al. 1999). Note that the standard definition of $B_T(\theta, \delta)$ is in terms of either the individual or the average expected relative entropy. For reasons that will become apparent later, we work with the total entropy, hence, to define information denseness we need to consider information balls of total entropy less or equal to $\delta_T T$ for any $\delta_T > 0$. Nevertheless, here we shall use a related and slightly weaker condition. To do so, we need to define the following quantity

$$R_T(\theta) := \inf_{\delta > 0} \{\delta - \ln w(B_T(\theta, \delta))\}$$

where $R_T(\theta)/T$ is called resolvability index (e.g. Barron, 1998). A candidate δ in the above display is of the form $\delta = \delta_T T$ where $\delta_T \to 0$ as $T \to 0$ (this is consistent with the notion of information denseness for neighbors of size $\delta_T T$). It can be shown that if w is information dense, then, $R_T(\theta)/T \to 0$ as $T \to \infty$ (Lemma 1). We state the condition that is used to show universality.

Condition 1

$$\lim_{T\to\infty} \sup_{\theta \in \Theta} \frac{R_T(\theta)}{T} = 0.$$

Information denseness and Condition 1 are slightly stronger than needed. In fact the following weaker condition would suffice: there is a set $A_T := A_T(\theta, \delta_T T) \subseteq \Theta$ such that

$$\mathbb{E}^{\theta} \ln p_{\theta}\left(Z_1^T\right) \leq \mathbb{E}^{\theta} \ln \left(\int_{A_T} p_{\theta'}\left(Z_1^T\right) \frac{w(d\theta')}{w(A_T)}\right) + \delta_T T \tag{5}$$

and $\{\delta_T T - \ln w(A_T)\}/T \to 0$ as $T \to \infty$. This clearly resembles the index of resolvability and requires $\delta_T \to 0$. It turns out that the set $B_T(\theta, \delta) \subseteq A_T(\theta, \delta)$ for any $\delta > 0$.

The following summarizes the above remarks.

Lemma 1 *An information dense prior w (at θ) implies $\lim_{T\to\infty} R_T(\theta)/T = 0$ and the latter implies (5) with $\lim_{T\to\infty}\{\delta_T T - \ln w(A_T)\}/T = 0$.*

In practice, verification of the above conditions is almost equivalent. Given that the index of resolvability provides an upper bound in most of the results, we shall use this as our default condition. Moreover, for two of the results to be stated (Theorem 5 and 6), (5) will not be sufficient. This suggests that Condition 1 is the relevant assumption to make for universality in a general framework.

By direct inspection of (4), Condition 1 is automatically satisfied with $\delta = 0$ if Θ is countable and finite and w puts strictly positive mass to each element of Θ (see the proof of Theorem 3, for details). Section 3.1 provides remarks on how to check Condition 1 in a special important case. The next section gives a fairly complete picture of universality of Bayesian predictions in a variety of contexts.

2 Universality Results

The previous section provided essential background on Bayesian prediction, its interpretations and discussed information denseness and negligibility of the resolvability index (Condition 1). Here we shall discuss universality results that can be derived from Condition 1 and obvious extensions to cover more general cases. At first, the standard well known result about Bayesian predictions is stated. Then, we show how this result can be used to prove Bayesian prediction under almost arbitrary loss functions. Furthermore, we look at universal bounds for Bayesian model averaging and the problem of Bayesian prediction in a non-stationary environment is discussed. In the last case, the standard posterior update is not adequate, but we can shrink the posterior in order to account for the uncertainty due to non-stationarity. Finally we discuss the problem of miss-specification. Explicit finite sample upper bounds are provided for most of these problems.

2.1 Universality of Probability Forecasts

The following establishes universality of Bayesian predictions in the simplest case.

Theorem 1 *Using the notation in (4)*

$$\sup_{\theta\in\Theta}\mathbb{E}^{\theta}D_{1,T}\left(P_{\theta}\|P_{w}\right)\leq\sup_{\theta\in\Theta}\inf_{\delta>0}\left\{\delta-\ln w\left(B_{T}\left(\theta,\delta\right)\right)\right\}$$

so that under Condition 1, the prediction is universal, i.e.

$$\sup_{\theta\in\Theta}\frac{1}{T}\mathbb{E}^{\theta}D_{1,T}\left(P_{\theta}\|P_{w}\right)\rightarrow0.$$

The upper bound is derived under no assumptions on the prior w and the r.h.s. can be infinite. Condition 1 makes sure that the bound is $o\left(T\right)$ as $T\rightarrow\infty$. Theorem 1 is well known (e.g. Barron, 1998) and it is a starting point for many other results to be discussed next. However, to give a simple econometric application of this result, consider the autoregressive process

$$Z_{t}=\theta Z_{t-1}+X_{t}$$

where $(X_t)_{t\in\mathbb{N}}$ is an iid sequence with distribution function $P(x)$ so that $P_\theta(z|\mathcal{F}_{t-1}) = P(z - \theta Z_{t-1})$, and $Z_0 = z$ is given. If $[0,1] \subseteq \Theta$, under Condition 1, we obtain universality even when $\theta = 1$, i.e. the Bayesian prediction performs uniformly well without need to worry about the possible presence of a unit root, and Theorem 1 gives a finite sample upperbound for the loss in the prediction. For example, in the Holder continuity case to be discussed in (15) (e.g. X_t is Gaussian noise, Cauchy, etc.), the resolvability index would be $O(\ln T/T)$. It is clearly unthinkable to derive such uniform finite sample upperbound in a maximum likelihood framework. We now turn to other related problems and defer any further discussion to Section 3.

2.2 Universal Predictions for Arbitrary Loss Functions

Suppose that $(Z_t)_{t\in\mathbb{N}}$ is a sequence of random variables with values in $\mathcal{Z}$. The problem is to find a prediction $f \in \mathfrak{F}$ for Z_{t+1}, where $\mathfrak{F}$ is a prespecified set. The framework is as follows: observe $Z_1, ..., Z_t$ and issue the prediction $f_{t+1} \in \mathfrak{F}$. Finally, Z_{t+1} is revealed and a loss $\mathcal{L}(Z_{t+1}, f_{t+1})$ is incurred, where the loss takes values in $\mathbb{R}_+$ (the non-negative reals). Our ideal goal is to minimize $\mathbb{E}_t^\theta \mathcal{L}(Z_{t+1}, f)$ w.r.t. $f \in \mathfrak{F}$, i.e. to find

$$f_{t+1}(\theta) := \arg\inf_{f\in\mathfrak{F}} \mathbb{E}_t^\theta \mathcal{L}(Z_{t+1}, f). \tag{6}$$

As in the previous section, we suppose that we only know the class $\{P_\theta : \theta \in \Theta\}$, but not under which θ expectation is taken. Hence, the problem is the one of finding a prediction that performs well for any $\theta \in \Theta$ and the given loss function. By suitable definition of $\mathcal{Z}$ and $\mathcal{L}$, the framework allows extra explanatory variables on top of autoregressive variables.

Example 2 *Suppose that $Z_t := (Y_t, X_t)$ and $\mathcal{Z} = \mathbb{R} \times \mathbb{R}$, and*

$$\mathcal{L}(Z_{t+1}, f) = |Y_{t+1} - f|^2.$$

Then, this is the usual problem of forecasting under the square loss using an autoregressive process plus an explanatory variable. In fact, if $P_\theta(\bullet|\mathcal{F}_t) = P_\theta(\bullet|Y_t, X_t)$ is Gaussian with mean $\theta_y Y_t + \theta_x X_t$ and finite variance, then,

$$\begin{aligned} f_{t+1}(\theta) &= \theta_y Y_t + \theta_x X_t \\ &= \arg\inf_{f\in\mathbb{R}} \mathbb{E}_t^\theta |Y_{t+1} - f|^2. \end{aligned}$$

Since θ is unknown, in (6) we shall replace the expectation w.r.t. $P_\theta(\bullet|\mathcal{F}_t)$ with expectation w.r.t. $P_w(\bullet|\mathcal{F}_t)$. This leads to the following prediction

$$f_{t+1}(w) := \arg\inf_{f\in\mathfrak{F}} \mathbb{E}_t^w \mathcal{L}(Z_{t+1}, f) \tag{7}$$

where $\mathbb{E}_t^w$ stands for expectation with respect to $P_w(\bullet|\mathcal{F}_t)$. We shall see that this prediction satisfies some desirable properties. To be more specific, we need the following.

Definition 2 *Predictions $f_1, ..., f_T$ are universal under $\mathcal{L}$ for $\{P_\theta : \theta \in \Theta\}$ if*

$$\sup_{\theta\in\Theta}\mathbb{E}^\theta \frac{1}{T}\sum_{t=1}^T \mathbb{E}_{t-1}^\theta \left[\mathcal{L}(Z_t, f_t) - \mathcal{L}(Z_t, f_t(\theta))\right] \to 0$$

as $T \to \infty$.

Remark 1 *As for the relative entropy, $\mathbb{E}_{t-1}^\theta\left[\mathcal{L}(Z_t, f_t) - \mathcal{L}(Z_t, f_t(\theta))\right] \geq 0$ by construction, because $f_t(\theta)$ is the predictor that minimizes the loss $\mathcal{L}$ under expectation w.r.t. $P_\theta(\bullet|\mathcal{F}_{t-1})$. Hence, universality implies*

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{t-1}^\theta \left[\mathcal{L}(Z_t, f_t) - \mathcal{L}(Z_t, f_t(\theta))\right] \to 0$$

in $L_1(P_\theta)$ and consequently in P_θ-probability for any $\theta \in \Theta$.

The following gives conditions under which the predictions $f_1(w), ..., f_T(w)$ are universal for a loss function $\mathcal{L}$.

Condition 2 *For any $\theta \in \Theta$ and $t \in \mathbb{N}$,*

$$\mathbb{E}^\theta\left[\mathbb{E}_{t-1}^\theta \mathcal{L}(Z_t, f_t(w))^r + \mathbb{E}_{t-1}^w \mathcal{L}(Z_t, f_t(\theta))^r\right] < \infty$$

for some $r > 1$.

Remark 2 *Further remarks on Condition 2 can be found in Section 4.2.*

We have the following result.

Theorem 2 *Under Condition 2,*

$$\sup_{\theta\in\Theta}\mathbb{E}^\theta \frac{1}{T}\sum_{t=1}^T \mathbb{E}_{t-1}^\theta \left[\mathcal{L}\left(Z_t, \hat{f}_t(w)\right) - \mathcal{L}(Z_t, f_t(\theta))\right] = o\left(\left[\frac{\sup_{\theta\in\Theta}\inf_{\delta>0}\{\delta - \ln w(B_T(\theta,\delta))\}}{T}\right]^{(r-1)/2r}\right)$$

and, if Condition 1 holds as well, the Bayesian predictions $f_1(w), ..., f_T(w)$ are universal.

Remark 3 *Theorem 2 says that if we use the Bayesian predictor (7), we can expect an average conditional prediction error asymptotically equal (in $L_1(P_\theta)$) to the average conditional prediction error obtained using the optimal predictions $f_1(\theta), ..., f_T(\theta)$. It is actually possible to write a proper upperbound in terms of constants that depend on the moments of the loss function only. In the case of a bounded loss function the rate of convergence is the square root of the one given by Theorem 1 up to a multiplicative constant (see the proof of Theorem 2 for details) .*

Merhav and Feder (1998) show how to relate the left hand side of Theorem 2 to the relative entropy in the case of bounded loss functions (by an application of Pinsker's inequality, e.g. Pollard 2002, eq. 13, p. 62). (See also Hutter, 2005, ch.3, for related results for bounded losses). The present result relates the expected difference of the loss functions to the resolvability index in the more general case of unbounded loss.

2.3 Universality of Bayesian Model Averaging

Parameter uncertainty in the model $\{P_\theta : \theta \in \Theta\}$ can be extended to model uncertainty. It is convenient to suppose K parameter spaces $\Theta_1, ..., \Theta_K$ within which each model is indexed, e.g. $\{P_\theta : \theta \in \Theta_k\}$ is model k. We shall define $\mathcal{K} := \{1, ..., K\}$. The Bayesian forecast of P_θ where $\theta \in \bigcup_{k \in \mathcal{K}} \Theta_k$ is given by

$$p_m(Z_t) := \sum_{k \in \mathcal{K}} p_{w_k}(Z_t | \mathcal{F}_{t-1}) \, m(k | \mathcal{F}_{t-1})$$

where

$$m(k | \mathcal{F}_t) = \frac{p_{w_k}(Z_t | \mathcal{F}_{t-1}) \, m(k | \mathcal{F}_{t-1})}{\sum_{k \in \mathcal{K}} p_{w_k}(Z_t | \mathcal{F}_{t-1}) \, m(k | \mathcal{F}_{t-1})}$$

$$p_{w_k}(z_t | \mathcal{F}_{t-1}) := \int_{\Theta_k} p_\theta(z_t | \mathcal{F}_{t-1}) \, dw_k(\theta | \mathcal{F}_{t-1})$$

and w_k, m are probability measures on subsets of Θ_k and $\mathcal{K}$, respectively. By induction, we have

$$p_m\left(Z_1^t\right) := \sum_{k \in \mathcal{K}} p_{w_k}\left(Z_1^t\right) m(k).$$

In this case, universality of the Bayesian prediction is understood as in Definition 1 where $\Theta := \bigcup_{k \in \mathcal{K}} \Theta_k$.

For universality we need the following additional condition.

Condition 3 *For any $k \in \mathcal{K}$, $m(k)$ is bounded away from zero.*

Hence, we can state the following.

Theorem 3 *We have the following upperbound,*

$$\max_{k \in \mathcal{K}} \sup_{\theta \in \Theta_k} \mathbb{E}^\theta D_{1,T}(P_\theta \| P_m) \leq \max_{k \in \mathcal{K}} \sup_{\theta \in \Theta_k} \inf_{\delta > 0} \{\delta - \ln w(B_T(\theta, \delta)) - \ln m(k)\},$$

so that under Condition 1 and 3, the predictions are universal, i.e.

$$\max_{k \in \mathcal{K}} \sup_{\theta \in \Theta_k} \frac{\mathbb{E}^\theta D_{1,T}(P_\theta \| P_m)}{T} \to 0.$$

Remark 4 *Condition 3 implies that $\mathcal{K}$ has finite cardinality. If $\mathcal{K}$ does not have finite cardinality, but the models are not too far away such that a condition equivalent to Condition 1 holds, then we* *still have universality. Details are exactly as in Theorem 1.*

The stated version of the upper bound is related to results derived in the machine learning and information theory literature (e.g. Cesa-Bianchi and Lugosi, 2006, and Sancetta, 2007 , for similar results in econometrics). The above references derive bounds for worst case scenarios and treat individual predictions to be combined as exogenous. The above bound also relates to some results in Yang, 2004, which apply to conditional mean prediction under the square loss.

2.4 Universality over Time Varying Reference Classes

In some situations we would like the Bayesian prediction to perform well when θ varies over time. We may think of this problem as the one when there are switches in regimes but we try not to make any assumptions on the dynamics (see Hamilton, 2005, for a review of parametric regime switches models). In this case, standard learning by Bayes rule is not appropriate and need to be modified. In fact, the application of Bayes theorem to derive P_w is based on θ constant overtime, i.e. it uses the joint distribution

$$P_\theta\left(Z_1^T\right) = \prod_{t=1}^{T} P_\theta\left(Z_t|\mathcal{F}_{t-1}\right)$$

while, here, we are interested in the joint distribution

$$P_{\theta_1^S}\left(Z_1^t\right) = \prod_{s=1}^{S} \prod_{t=T_{s-1}+1}^{T_s} P_{\theta_s}\left(Z_t|\mathcal{F}_{t-1}\right) \tag{8}$$

where $\theta_1^S := (\theta_1, ..., \theta_S)$, and $0 = T_0 < T_1 < ... < T_s = T$ are arbitrary, but fixed.

Example 3 *Suppose that $P_{\theta_s}\left(Z_s|\mathcal{F}_{s-1}\right) = P_{\theta_s}\left(Z_s|Z_{s-1} = z_{s-1}\right)$ is a Markov transition distribution. If θ_s does not vary over time, the transition distribution is homogeneous (i.e. stationary). Allowing for θ_s to vary with time leads to a inhomogeneous Markov transition distribution.*

To ease notation define the time segments $\mathcal{T}_s := (T_{s-1}, T_s] \cap \mathbb{N}$. For $s \leq S$, we shall denote expectation w.r.t. $P_{\theta_1^s}$ by $\mathbb{E}^{\theta_1^s}$. To be precise, the notation should make explicit not only θ_1^s, but also $\mathcal{T}_1, ..., \mathcal{T}_S$. For simplicity the times of the parameter's change are omitted, as they will be clear from the context, if necessary.

The problem of universality of the predictions is formalized by the following definition.

Definition 3 *The prediction p_w is universal for $\left\{P_{\theta_1^S} : \theta_1^S \in \Theta^S\right\}$ over $S \leq T$ partitions if*

$$\max_{\mathcal{T}_1,...,\mathcal{T}_S} \frac{1}{T} \sup_{\theta_1^S \in \Theta^S} \mathbb{E}^{\theta_1^S} \sum_{s=1}^{S} \sum_{t \in \mathcal{T}_s} D_t\left(P_{\theta_s} \| P_w\right) \to 0$$

as $T \to \infty$.

Note that in the above definition S may go to infinity with T. To allow for changing θ when the time of change is not known apriori, we need to introduce a prior on the probability of changes. The simplest approach that leads to constructive results is to define a probability measure on subsets of $\mathbb{N}$: for each t, $\lambda_t(r)$ is a probability density w.r.t. the counting measure with support in $\{0,2,...,t\}$, so that $\sum_{r=0}^{t}\lambda_t(t-r)=1$. Then we mix past posteriors using $\lambda_t(r)$ as mixing density:

$$w(d\theta|\mathcal{F}_t)=\sum_{r=0}^{t}\lambda_t(t-r)\,w'(d\theta|\mathcal{F}_{t-r}) \tag{9}$$

where

$$w'(d\theta|\mathcal{F}_0)=w(d\theta|\mathcal{F}_0)$$

and

$$w'(d\theta|\mathcal{F}_t)=\frac{p_\theta(Z_t|\mathcal{F}_{t-1})\,w(d\theta|\mathcal{F}_{t-1})}{\int_\Theta p_\theta(Z_t|\mathcal{F}_{t-1})\,w(d\theta|\mathcal{F}_{t-1})}. \tag{10}$$

The Bayesian interpretation is that with probability $\lambda_t(r)$ the posterior of θ at time t is equal to the posterior $dw'(\theta|\mathcal{F}_r)$ at time $r+1<t$. This means that at any point in time we may expect shifts that take us back to a past regime. When $r=0$ we are taken back to the prior, which corresponds to the start of a new regime that has not previously occurred. This is the intuition behind (9) and will be further developed next.

We shall use $D_{\mathcal{T}_s}(P_\theta\|P_{\theta'}):=D_{T_{s-1}+1,T_s}(P_\theta\|P_{\theta'})$ for the relative entropy over the time interval $\mathcal{T}_s$. To prove universality, we need a condition slightly stronger than Condition 1.

Condition 4 *For any $\theta_s\in\Theta$, $\mathcal{T}_s$, $s\leq S$ and $\delta>0$ define the following set*

$$B_{\mathcal{T}_s}(\theta_s,\delta):=\left\{\theta'\in\Theta:\mathbb{E}^{\theta_1^s}D_{\mathcal{T}_s}(P_{\theta_s}\|P_{\theta'})\leq\delta\right\}$$

and the following unstandardized resolvability index

$$R_{\mathcal{T}_s}(\theta_s):=\inf_{\delta_s>0}\left[\delta_s-\ln w\left(B_{\mathcal{T}_s}(\theta_s,\delta_s)\right)\right]$$

Then,

$$\lim_{T\to\infty}\sup_{\theta_1^S\in\Theta^S}\sum_{s=1}^{S}\frac{R_{\mathcal{T}_s}(\theta_s)}{T}=0.$$

For definiteness, two special cases will be considered. In one case we make no assumption on the type of changes, and only assume that there are $S-1$ changes. Hence, in this case any change could be a new regime and past information might be useless. For this reason, we shall just shrink the posterior towards the prior. In the second case, we assume that there are $S-1$ shifts in the parameter, but that these shifts are back and forth within a small number of $V<S$ regimes (i.e. parameters). The details will become clear in due course.

2.4.1 Shrinking towards the Prior

We restrict λ_t such that $\lambda_t(t) = 1 - \lambda t^{-\alpha}$, $\lambda_t(0) = \lambda t^{-\alpha}$, and $\lambda_t(r) = 0$ otherwise, with $\alpha \geq 0$ and $\lambda \in (0,1)$. This means that (9) simplifies to

$$w(d\theta|\mathcal{F}_t) = \left(1 - \lambda t^{-\alpha}\right) w'(d\theta|\mathcal{F}_t) + \lambda t^{-\alpha} w(d\theta). \tag{11}$$

Theorem 4 *Using (11), for any segments $\mathcal{T}_1, ..., \mathcal{T}_S$,*

$$\begin{aligned} &\sup_{\theta_1^S \in \Theta^S} \mathbb{E}^{\theta_1^S} \sum_{s=1}^{S} \sum_{t \in \mathcal{T}_s} D_t\left(P_{\theta_s} \| P_w\right) \\ \leq\ & \sup_{\theta_1^S \in \Theta^S} \sum_{s=1}^{S} \inf_{\delta_s > 0} \left[\delta_s - \ln w\left(B_{\mathcal{T}_s}(\theta_s, \delta_s)\right)\right] \\ &+ \frac{2\lambda}{\sqrt{1-\lambda^2}} \left(1 + \frac{T^{1-\alpha} - 1}{1-\alpha}\right) + S \ln(1/\lambda) + \alpha S \ln T \end{aligned}$$

so that the prediction is universal under Condition 4 if $S \ln T = o(T)$.

Remark 5 *If $\alpha \to 1$, $\left(T^{1-\alpha} - 1\right)/(1-\alpha) \to \ln T$; in fact, the second term in the bound of Theorem 4 is monotonically decreasing in α. Increasing α does however increase the last term in the bound, i.e. $\alpha S \ln T$.*

In the bound of Theorem 4, α and λ are free parameters whose choice can be based on prior knowledge or subjective believes. If S is of large order, we could minimize the bound setting λ close to one and α close to zero. This is just a loose remark whose only purpose is to suggest that as the number of shifts increases relatively to T, we are better off shrinking towards the prior. This idea can be related to the debate about equally weighted model averaging when we want to hedge against non-stationarity (e.g. Timmermann, 2006, for discussions). Clearly, exact prior knowledge of T (in the sense of number of predictions to be made) and S would allow us to minimize the bound w.r.t. the free parameters.

In Theorem 4,

$$\sup_{\theta_1^S \in \Theta^S} \frac{1}{T} \sum_{s=1}^{S} \inf_{\delta_s > 0} \left[\delta_s - \ln w\left(B_{\mathcal{T}_s}(\theta_s, \delta_s)\right)\right] = o(1)$$

by Condition 4. However the above resolvability index can be quite large as the order of magnitude of S increases. Moreover, all the shifts might not be to new regimes, hence, it could be advantageous to use past information hoping to reduce the resolvability index. This issue will be addressed next.

2.4.2 Improvements on the Resolvability Index: Switching within a Small Number of Parameters

We now consider the case of shifting parameter within a set of V fixed parameters. Hence, even if $S \to \infty$ we may still have $V = O(1)$ so that over the $S-1$ shifts we move back and forth V regimes. In particular, to setup notation, there are $S-1$ shifts within $\left\{\tilde{\theta}_1, ..., \tilde{\theta}_V\right\} \subset \Theta$, $V < S$. Hence, for given $\tilde{\theta}_v$, there are $S_v \leq \lfloor S/V \rfloor + 1$ segments of the kind $[T_{s-1}+1, T_s]$ for which $\theta_s = \tilde{\theta}_v$ is the "true parameter". By the intuition that using past information should be helpful, we may hope to improve on the bound of Theorem 4 letting $\lambda_t(r) > 0$ for any $r \leq t$. This is indeed the case and to this end we state the following.

Condition 5 *For any $\theta_s \in \Theta$, $\mathcal{T}_s$, $s \leq S$ and $\delta_1^S := (\delta_1, ..., \delta_S) > 0$ (understood elementwise), define the following set*

$$B_v\left(\tilde{\theta}_v, \delta_1^S\right) := \bigcap_{\left\{s:\theta_s = \tilde{\theta}_v\right\}} B_{\mathcal{T}_s}(\theta_s, \delta_s)$$

i.e. the smallest set $B_{\mathcal{T}_s}(\theta_s, \delta_s)$ w.r.t. s such that $\theta_s = \tilde{\theta}_v$, where $B_{\mathcal{T}_s}(\theta_s, \delta_s)$ is as in Condition 4. Then,

$$\lim_{T\to\infty} \sup_{\theta_1^S \in \Theta^S} \inf_{\delta_1^S > 0} \left\{\sum_{s=1}^{S} \delta_s - \sum_{v=1}^{V} \ln w\left(B_v\left(\tilde{\theta}_v, \delta_1^S\right)\right)\right\} = 0.$$

Remark 6 *Note that*

$$\ln w\left(B_v\left(\tilde{\theta}_v, \delta_1^S\right)\right) \leq \min_{\left\{s:\theta_s = \tilde{\theta}_v\right\}} \ln w\left(B_{\mathcal{T}_s}(\theta_s, \delta_s)\right)$$

with equality in some special important cases as in (15).

The simplest approach to let $\lambda_t(r) > 0$ for $r \in [0, t]$ is to directly extend the density $\lambda_t(r)$ in the previous subsection: $\lambda_t(t) = 1 - \lambda t^{-\alpha}$, $\lambda_t(r) = \lambda t^{-(1+\alpha)}$ when $r \in [0, t)$ and α and λ are as previously constrained. Direct calculation shows that $\lambda_t(r)$ is a probability density (w.r.t. the counting measure) on $[0, t] \cap \mathbb{N}$, leading to the following posterior update

$$w(d\theta|\mathcal{F}_t) = \left(1 - \lambda t^{-\alpha}\right) w'(d\theta|\mathcal{F}_t) + \sum_{r=1}^{t} \frac{\lambda t^{-\alpha}}{t} w'(d\theta|\mathcal{F}_{t-r}). \quad (12)$$

Under the above update, we can derive the following bound for $S-1$ shifts within V regimes.

Theorem 5 *Using (12), for any segments $\mathcal{T}_1, ..., \mathcal{T}_S$, for S shifts in θ_s within a fixed but arbitrary* *set* $\left\{\tilde{\theta}_1, ..., \tilde{\theta}_V\right\}$ *with* $V \leq S$,

$$\begin{aligned} & \sup_{\theta_1^S \in \left\{\tilde{\theta}_1,...,\tilde{\theta}_V\right\}^S} \mathbb{E}^{\theta_1^S} \sum_{s=1}^{S} \sum_{t \in \mathcal{T}_s} D_t \left(P_{\theta_s} \| P_w\right) \\ \leq \; & \inf_{\delta_1^S > 0} \left\{ \sum_{s=1}^{S} \delta_s - \sum_{v=1}^{V} \ln w \left(B_v \left(\tilde{\theta}_v, \delta_1^S\right)\right) \right\} \\ & + \frac{2\lambda}{\sqrt{1 - \lambda^2 S^{-2\alpha}}} \left(S^{-\alpha} + \frac{T^{1-\alpha} - S^{1-\alpha}}{1-\alpha}\right) + S \ln (1/\lambda) + (1+\alpha) S \ln T \end{aligned}$$

so that the prediction is universal under Condition 5 if $S \ln T = o(T)$.

Remark 7 *Theorem 5 leads to a considerable decrease in the resolvability index when* V *is fixed and* $S \to \infty$. *However, comparison with Theorem 4 shows that this comes at the extra cost of an error term* $S \ln T$ *together with an improvement in*

$$\frac{2\lambda}{\sqrt{1 - \lambda^2 S^{-2\alpha}}} \left(S^{-\alpha} + \frac{T^{1-\alpha} - S^{1-\alpha}}{1-\alpha}\right). \tag{13}$$

Section 3.3 provides further remarks on the improvement in the resolvability index using $\lambda_t(r) > 0$ *for* $r \in [0, t]$ *when there are only* V *regimes, in a special important case. For the case to be considered in Section 3.3, it can be shown that the gain in the resolvability index together with the gain in (13) is offset by* $S \ln T$*, though only asymptotically. It is a matter of simple algebra to show that for finite* T *and large* S *we can find* $\alpha \simeq 0$ *and* λ *close to one such that the result in Theorem 5 strictly improves Theorem 4. Moreover, for comparisons, we do not need the* α *in Theorem 5 to be the same as in Theorem 4. However, note that Theorems 4 and 5 only provide upperbounds, so that one has to be cautious about comparisons. When* Θ *is countable and finite, Bousquet and Warmuth (2002) provide encouraging simulation evidence in favor of mixing past posteriors using* $\lambda_t(r) > 0$ $(r \in [0, t])$ *when* V *is small and* S *is large. This is exactly the case when one would be expected to use* α *close to zero and* λ *close to one (recall the discussion just after Theorem 4). According to these remarks, the mixing update in (12) should be used with small* α *and large* λ *if we expect* S *to be relatively large and* V *small so that the resulting loss should dominate the one incurred using the update in (11).*

We now consider a second case that further improves on the previous result. This can be achieved by letting $\lambda_t(r)$ put less and less mass on the remote past. To this end we consider the following simple case: $\lambda_t(t) = 1 - \lambda t^{-\alpha}$, $\lambda_t(r) = \lambda t^{-\alpha} A_t^{-1} (1 + t - r)^{-2}$, for $0 \leq r < t$ where $A_t = \sum_{r=0}^{t-1} (1 + t - r)^{-2}$ is a normalizing factor and α and λ are as previously restricted. This means that we shall consider the following update

$$w(d\theta|\mathcal{F}_t) = \left(1 - \lambda t^{-\alpha}\right) w'(d\theta|\mathcal{F}_t) + \sum_{r=1}^{t} \frac{\lambda t^{-\alpha}}{A_t (1+r)^2} w'(d\theta|\mathcal{F}_{t-r}). \tag{14}$$

Theorem 6 *Using (14) instead of (12) in Theorem 5,*

$$
\begin{aligned}
&\sup_{\theta_1^S \in \left\{\tilde{\theta}_1,\ldots,\tilde{\theta}_V\right\}^S} \mathbb{E}^{\theta_1^S} \sum_{s=1}^{S} \sum_{t=T_{s-1}+1}^{T_s} D_t\left(P_{\theta_s} \| P_w\right) \\
\leq\ & \inf_{\delta_1^S>0} \left\{ \sum_{s=1}^{S} \delta_s - \sum_{v=1}^{V} \ln w\left(B_v\left(\tilde{\theta}_v\right)\right) \right\} + \frac{2\lambda}{\sqrt{1-\lambda^2 S^{-2\alpha}}} \left(S^{-\alpha} + \frac{T^{1-\alpha} - S^{1-\alpha}}{1-\alpha} \right) \\
& S \ln\left(1/\lambda\right) + \alpha S \ln T + 2S \ln\left(\frac{V\left(T-1\right)}{S-1}\right)
\end{aligned}
$$

so that the prediction is universal under Condition 5 if $S \ln T = o\left(T\right)$.

Remark 8 *Theorem 6 shows that the extra cost* $S \ln T$ *in Theorem 5 can be reduced to* $2S \ln\left(\frac{V(T-1)}{S-1}\right)$ *if we use (14) instead of (12).*

Mutatis mutandis, Theorem 4, 5 and 6 are related to Lemma 6 and Corollary 8 and 9 in Bousquet and Warmuth (2002) and improve on the bounds given by these authors using slightly different functions to mix posteriors. Bousquet and Warmuth (2002) were the first to propose predictions by mixing past posteriors (see also Herbster and Warmuth, 1998, for related results). They are essentially concerned with the forecast combination problem, called prediction with experts' advice in the machine learning literature. The main difference lies in the fact that they use a finite and countable parameter space, while here the parameter space is possibly uncountable, given the Bayesian prediction's setting. The machine learning literature is rich of results of this kind which can often be justified by Bayesian arguments.

By the same method of proof, we can consider other mixing distributions. For example, the case $\lambda_t\left(r\right) = \lambda t^{-\alpha} A_t^{-1} \left(1+t-r\right)^{-\gamma}$ $(r < t)$, where $\gamma > 2$, with suitably modified A_t, is dealt similarly, but seems to lead to a more complex bound.

2.5 Bounds when the True Model is not in the Reference Class

The previous results considered the case where expectation is taken with respect to one element within a class of models, e.g. $\{P_\theta : \theta \in \Theta\}$. This implies that we only face estimation error. However, when expectation is taken with respect to a probability $P \notin \{P_\theta : \theta \in \Theta\}$, we shall also incur an approximation error. This approximation error can be characterized in terms of the relative entropy. With no loss of generality, we assume that P is absolutely continuous w.r.t. the sigma finite measure μ and we denote its density by p, so that

$$D_t\left(P \| P_\theta\right) = \mathbb{E}_{t-1} \ln \frac{p\left(Z_t | \mathcal{F}_{t-1}\right)}{p_\theta\left(Z_t | \mathcal{F}_{t-1}\right)}$$

where $\mathbb{E}_{t-1}$ is expectation w.r.t. $P\left(\bullet | \mathcal{F}_{t-1}\right)$. Note that this does not imply that P is absolutely continuous w.r.t. P_θ, however, if this is not the case, their relative entropy is infinite. We shall also use $\mathbb{E}$ for (unconditional) expectation w.r.t. P. We need the following condition that extends Condition 2 to the present more general framework.

Condition 6 *Define*

$$f_t(P) := \arg\inf_{f \in \mathfrak{F}} \mathbb{E}_{t-1}\mathcal{L}(Z_t, f).$$

Then, for any $\theta \in \Theta$ and $t \in \mathbb{N}$,

$$\mathbb{E}\left[\mathbb{E}_{t-1}\mathcal{L}(Z_t, f_t(w))^r + \mathbb{E}^w_{t-1}\mathcal{L}(Z_t, f_t(P))^r\right] < \infty$$

for some $r > 1$.

Then, we have the following that also gives the extra error term due to the approximation.

Theorem 7 *Under Condition 6*

$$\begin{aligned} & \mathbb{E}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{t-1}\left[\mathcal{L}\left(Z_t, \hat{f}_t(w)\right) - \mathcal{L}(Z_t, f_t(P))\right] \\ = \quad & o\left(\left[\frac{\inf_{\theta\in\Theta}\inf_{\delta}\{\mathbb{E}D_{1,T}(P\|P_\theta) + \delta - \ln w(B_T(\theta,\delta))\}}{T}\right]^{(r-1)/2r}\right). \end{aligned}$$

Remark 9 *By the following inequality*

$$\begin{aligned} & \inf_{\theta\in\Theta}\inf_{\delta}\{\mathbb{E}D_{1,T}(P\|P_\theta) + \delta - \ln w(B_T(\theta,\delta))\} \\ \leq \quad & \inf_{\theta\in\Theta}\mathbb{E}D_{1,T}(P\|P_\theta) + \sup_{\theta\in\Theta}\inf_{\delta}\{\delta - \ln w(B_T(\theta,\delta))\} \end{aligned}$$

we deduce that if Condition 1 holds, the Bayesian prediction might not be universal, but will lead to the smallest possible information loss, i.e. $\inf_{\theta\in\Theta}\mathbb{E}D_{1,T}(P\|P_\theta)/T$.

3 Discussion

3.1 Remarks on Condition 1

Verification of Condition 1 requires smoothness of the total relative entropy. For simplicity suppose $\Theta \subset \mathbb{R}$ (the discussion easily extends to more general metric spaces, not just Euclidean spaces). Smoothness can be formalized in terms of a Holder's continuity condition: for any $t \in \mathbb{N}$

$$\mathbb{E}^\theta[\ln p_{\theta'}(Z_t|\mathcal{F}_{t-1}) - \ln p_\theta(Z_t|\mathcal{F}_{t-1})] \leq b|\theta' - \theta|^a \tag{15}$$

for some $a, b > 0$. In this case, we set $\delta = Tb|\theta' - \theta|^a$ and

$$B_T(\theta,\delta) = \left\{\theta' \in \Theta : |\theta' - \theta| \leq \left(\frac{\delta}{Tb}\right)^{1/a}\right\}.$$

Assuming for simplicity the Lebesgue measure as prior and Θ having unit Lebesgue measure, $w\left(B_T\left(\theta,\delta\right)\right)=\left[\delta/\left(Tb\right)\right]^{1/a}$. Then,

$$R_T\left(\theta\right)=\inf_{\delta>0}\left\{\delta-\frac{1}{a}\ln\left(\frac{\delta}{Tb}\right)\right\}$$

which is minimized by $\delta=a^{-1}$ so that the resolvability index is equal to

$$\frac{R_T\left(\theta\right)}{T}=\frac{1+\ln\left(abT\right)}{aT}$$

and the average relative entropy converges to zero at the rate $\ln T/T$ for any Holder's continuous class of expected conditional log-likelihoods. To put (15) into perspective, note that differentiability of the expected conditional log-likelihood per observation is stronger than (15). We give a prototypical example where standard maximum likelihood methods are known to fail for some parameter values.

Example 4 *Suppose $\left(Z_t\right)_{t\in\mathbb{N}}$ is a sequence of iid random variables with double exponential density $p_\theta\left(z\right)=2^{-1}\exp\left\{-\left|z-\theta\right|\right\}$. Then, (15) holds with $a=1$, while p_θ is not differentiable at $\theta=0$.*

3.2 Remarks on Condition 2

Condition 2 needs to be checked on a case by case basis and might be hard to verify except for some special cases (e.g. when $\mathcal{L}$ is the square loss and p_θ is Gaussian). Simplicity can be gained by restricting the set $\mathfrak{F}$ over which to carry out minimization. For example, we may choose $\mathfrak{F}$ to contain all the function such that $|f|\leq g$ where g is some measurable function such that $\sup_{\theta\in\Theta}\mathbb{E}^\theta g<\infty$. In this case, restrictions on the loss function may lead to feasible computations. We provide a simple example next.

Example 5 *Suppose $p_\theta\left(Z_t|\mathcal{F}_{t-1}\right)=p_\theta\left(Z_t|Z_{t-1}\right)$ is a Markov transition density. Then, we may restrict $\mathfrak{F}$ to contain only functions f such that $\left|f\left(z\right)\right|\leq g\left(z\right)=1+b\left|z\right|^a$ for some $a,b>0$. Suppose that the loss function can be bounded as follows $\mathcal{L}\left(z,f\right)\leq|z|+|f|$. Then, to check Condition 2 it is sufficient to check*

$$\mathbb{E}^\theta\mathcal{L}\left(Z_t,f_t\left(w\right)\right)^r+\mathbb{E}^\theta\mathbb{E}^w_{t-1}\mathcal{L}\left(Z_t,f_t\left(\theta\right)\right)^r\lesssim\mathbb{E}^\theta\left(\mathbb{E}^\theta_{t-1}+\mathbb{E}^w_{t-1}\right)\left|Z_t\right|^r+\mathbb{E}^\theta\left|Z_{t-1}\right|^{ar}$$

and the right hand bound might be easier to deal with ($\lesssim$ is $\leq$ up to a multiplicative finite absolute constant).

.

3.3 Improvement on the Resolvability Index of Theorem 6 over Theorem 4

Consider the Holder's continuity condition in (15) and the same prior as given in its discussion. To simplify suppose that all the time segments $\mathcal{T}_s$ have same length $T/S \in \mathbb{N}$. Then we shall choose

$$B_{\mathcal{T}_s}(\theta_s, \delta) = \left\{ \theta' \in \Theta : |\theta' - \theta| \leq \left(\frac{S\delta}{Tb} \right)^{1/a} \right\}$$

implying in Theorem 4

$$\begin{aligned} \sum_{s=1}^{S} \inf_{\delta_s > 0} \{ \delta_s - \ln w (B_{\mathcal{T}_s}(\theta_s, \delta_s)) \} &= S \inf_{\delta > 0} \left\{ \delta - \frac{1}{a} \ln \left(\frac{S\delta}{Tb} \right) \right\} \\ &= \frac{S}{a} \left(1 + \ln \frac{Tab}{S} \right) \end{aligned}$$

substituting the minimizer $\delta = a^{-1}$. Clearly, if S is of large order this quantity will be large. On the other hand, in Theorem 6 we would have

$$\begin{aligned} \inf_{\delta_1^S > 0} \left\{ \sum_{s=1}^{S} \delta_s - \sum_{v=1}^{V} \ln w \left(B_v \left(\tilde{\theta}_v, \delta_1^S \right) \right) \right\} &= \inf_{\delta > 0} \left\{ S\delta - V \frac{1}{a} \ln \left(\frac{S\delta}{Tb} \right) \right\} \\ &= \frac{V}{a} \left\{ 1 + \ln \frac{abT}{V} \right\} \end{aligned}$$

substituting the minimizer $\delta = V/(aS)$. Unlike the former, this latter bound does not depend on the number of shifts S.

3.4 Further Remarks

This paper provided a comprehensive set of results for universal prediction using Bayes rule. The conditions used restricted Θ only implicitly. For Condition 1 to hold, Θ cannot be completely arbitrary, but the restrictions on Θ are quite mild. In fact, we could let Θ be a set of densities and w a prior on it. Hence, the results stated here are not necessarily restricted to parametric models (e.g. Barron et al, 1999, for results in this direction).

The relative improvement on the resolvability index when we mix past posteriors (and not just the prior, i.e. (11)) might be offset by an extra term that enters the error bound. This extra term depends on the mixing update. For the updates considered, it is possible to show superiority in finite samples only in some special cases by fine tuning of α and λ. Given that the improvement on the resolvability index is independent of the mixing scheme (as long as $\lambda_t(r) > 0$ for $r \in [0, t]$) one could try to study and compare different updates. For example, we showed that (14) already improved upon (12). Perhaps, more definite claims could be made if a different method of proof were used.

There is a number of topics of practical relevance that have not been discussed. Among the most important omitted issues are computational issues, but references have been provided in the Introduction. In general, computational improvements may be obtained by restricting Θ to be compact and choose a prior from which simulation is easy. Computational problems in Bayesian methods is an active area of research.

Some theoretical issues not discussed here deserve attention. In particular the problem of model complexity should be mentioned. An implicit measure of model complexity is given by Condition 1 and related ones. There are links between the Bayesian information criterion and other measures of complexity like the minimum description length principle of Rissanen (e.g. Rissanen, 1986, Barron et al., 1998). The relation between complexity (in a computable sense) and prior distribution has also been discussed in the artificial intelligence literature (Hutter, 2005, for details). Tight estimates of model complexity are the key for tight and explicit rates of convergence of Bayesian predictions.

Another issue not discussed is the multiple steps ahead prediction problem, where we want to use Z_1^t to make (distributional) predictions about Z_{t+h}, for fixed $h > 1$. Unfortunately, it seems that the relative entropy is too strong to derive bounds in this case, while results can be easily derived using the total variation distance (Hutter, 2005, sect. 3.7.1, for illustrations when $\mathcal{Z}$ is countable). To the author's knowledge this is an open problem. Nevertheless, bounds under the relative entropy for distributional prediction of Z_t^{t+h} given Z_1^{t-1} can be derived directly from the results given in this paper. Just note that, in this case, the relative entropy is given by

$$\mathbb{E}_{t-1}^{\theta} \ln \frac{p_\theta\left(Z_t^{t+h}|\mathcal{F}_{t-1}\right)}{p_w\left(Z_t^{t+h}|\mathcal{F}_{t-1}\right)} = \mathbb{E}_{t-1}^{\theta} \ln \frac{p_\theta\left(Z_1^{t+h}\right)}{p_w\left(Z_1^{t+h}\right)} - \mathbb{E}_{t-1}^{\theta} \ln \left[\frac{p_\theta\left(Z_1^{t-1}\right)}{p_w\left(Z_1^{t-1}\right)}\right] \{t > 1\} \tag{16}$$

using (1) (see Lemma 2 for the derivation). Hence, summing over t and taking full expectation, the sum telescopes apart from initial h negative terms which can be disregarded in the upper bound plus the last $h+1$ terms which are kept:

$$\begin{aligned}
\mathbb{E}^{\theta} \sum_{t=1}^{T} \mathbb{E}_{t-1}^{\theta} \ln \frac{p_\theta\left(Z_t^{t+h}|\mathcal{F}_{t-1}\right)}{p_w\left(Z_t^{t+h}|\mathcal{F}_{t-1}\right)} &\leq \sum_{t=T}^{T+h} \mathbb{E}^{\theta} \mathbb{E}_{t-1}^{\theta} \ln \frac{p_\theta\left(Z_1^{t+h}\right)}{p_w\left(Z_1^{t+h}\right)} \\
&\leq (h+1)\, \mathbb{E}^{\theta} \ln \frac{p_\theta\left(Z_1^{T+h}\right)}{p_w\left(Z_1^{T+h}\right)} \\
&\quad [\text{the joint KL divergence is increasing in } T] \\
&= (h+1)\, D_{1,T}\left(P_\theta \| P_w\right).
\end{aligned}$$

The above display shows that the bounds grow linearly in h. In order to derive an h steps ahead prediction we could start from the joint conditional distribution of Z_t^{t+h} and integrate out Z_t^{t+h-1}. Unfortunately, doing so, (16) is not valid anymore. Moreover, the above approach does not allow us to work directly with the h steps ahead predictive distribution and requires specifying the joint distribution of a segment given the past, which is potentially a more difficult task. More research effort is required in this direction using possibly different convergence requirements.

A Appendix: Proofs

The proofs may refer to some technical lemmata stated at the end of the section.

Proof. [Lemma 1] Information denseness implies $-\ln w\left(B_T\left(\theta,\delta_T T\right)\right)<\infty$ for any $\delta_T>0$. Hence $\delta_T-T^{-1}\ln w\left(B_T\left(\theta,\delta_T T\right)\right)$ can be made arbitrary small by choosing $\delta_T\rightarrow 0$. This implies $R_T\left(\theta\right)/T\rightarrow 0$. To show the last implication, define

$$p_{w,A_T}\left(z_1^T\right):=\int_{A_T(\theta)}p_{\theta'}\left(z_1^T\right)\frac{w\left(d\theta'\right)}{w\left(A_T\left(\theta\right)\right)}$$

for $A_T\left(\theta\right):=A_T\left(\theta,\delta_T T\right)$ such that

$$D_{1,T}\left(P_\theta\|P_{w,A_T}\right)\leq\delta_T T \tag{17}$$

which is (5). Setting $B_T\left(\theta\right):=B_T\left(\theta,\delta_T T\right)$,

$$\begin{aligned}D_{1,T}\left(P_\theta\|P_{w,B_T}\right) &\leq \int_{B_T(\theta)}\mathbb{E}^\theta\ln\left(\frac{p_\theta\left(Z_1^T\right)}{p_{\theta'}\left(Z_1^T\right)}\right)\frac{w\left(d\theta'\right)}{w\left(B_T\left(\theta\right)\right)}\\ &\quad\text{[by Jensen's inequality]}\\ &\leq \sup_{\theta'\in B_T(\theta)}\mathbb{E}^\theta\ln\left(\frac{p_\theta\left(Z_1^T\right)}{p_{\theta'}\left(Z_1^T\right)}\right)\\ &\leq \delta_T T\end{aligned}$$

by definition of $B_T\left(\theta\right)$. The above inequality together with (17) imply that $B_T\left(\theta,\delta_T T\right)\subseteq A_T\left(\theta,\delta_T T\right)$. ■

Proof. [Theorem 1] Choosing a ball $B\left(\theta\right):=B_T\left(\theta\right)$ as in (4),

$$\begin{aligned}\mathbb{E}^\theta\ln\int_\Theta p_{\theta'}\left(Z_1^T\right)w\left(d\theta'\right) &\geq \mathbb{E}^\theta\ln\int_{B(\theta)}p_{\theta'}\left(Z_1^T\right)w\left(d\theta'\right)\\ &\quad\text{[because } p_\theta\left(Z_1^T\right)\text{ is non-negative]}\\ &\geq \mathbb{E}^\theta\ln\left(p_\theta\left(Z_1^t\right)\right)-\delta+\ln w\left(B\left(\theta\right)\right) \end{aligned}\tag{18}$$

by the same arguments as in the proof of Lemma 1 noting that

$$\ln\int_{B(\theta)}p_{\theta'}\left(Z_1^T\right)w\left(d\theta'\right)=\ln\int_{B(\theta)}p_{\theta'}\left(Z_1^T\right)\frac{w\left(d\theta'\right)}{w\left(B\left(\theta\right)\right)}+\ln w\left(B\left(\theta\right)\right).$$

Hence,

$$
\begin{aligned}
\mathbb{E}^{\theta} D_{1,T}\left(P_{\theta} \| P_{w}\right) &= \mathbb{E}^{\theta} \sum_{t=1}^{T} \mathbb{E}_{t-1}^{\theta}\left[\ln\left(p_{\theta}\left(Z_t | \mathcal{F}_{t-1}\right)\right) - \ln\left(p_{w}\left(Z_t | \mathcal{F}_{t-1}\right)\right)\right] \\
&= \mathbb{E}^{\theta}\left[\ln p_{\theta}\left(Z_1^T\right) - \ln p_{w}\left(Z_1^T\right)\right] \\
&\qquad [\text{because } \mathcal{F}_0 \text{ is trivial, using Lemma 2}] \\
&\leq \delta - \ln w\left(B\left(\theta\right)\right)
\end{aligned}
$$

by (18). Given that the above bound holds for any $\delta > 0$ (with the r.h.s. possibly infinite) we can take $\sup_{\theta \in \Theta} \inf_{\delta}$ on both sides and obtain the result. ■

Notation 1 *If A is a set, we directly use A in place of its indicator function I_A.*

Proof. [Theorem 2] Define $\Delta_t\left(w, \theta\right) := \mathcal{L}\left(Z_t, f_t\left(w\right)\right) - \mathcal{L}\left(Z_t, f_t\left(\theta\right)\right)$. Then $\mathbb{E}_{t-1}^{w} \Delta_t\left(w, \theta\right) \leq 0$ because $f_t\left(w\right)$ is the minimizer of $\mathbb{E}_{t-1}^{w} \mathcal{L}\left(Z_t, f\right)$. Define the sets $M_w := \left\{\mathcal{L}\left(Z_t, f_t\left(w\right)\right) \leq M\right\}$ and $M_{\theta} := \left\{\mathcal{L}\left(Z_t, f_t\left(\theta\right)\right) \leq M\right\}$ and denote their complements by M_w^c and M_{θ}^c. By this remark, adding and subtracting $\mathbb{E}_{t-1}^{w} \Delta_t\left(w, \theta\right)$,

$$
\begin{aligned}
\mathbb{E}_{t-1}^{\theta} \Delta_t\left(w, \theta\right) &= \mathbb{E}_{t-1}^{w} \Delta_t\left(w, \theta\right) + \left(\mathbb{E}_{t-1}^{\theta} - \mathbb{E}_{t-1}^{w}\right) \Delta_t\left(w, \theta\right) \\
&\leq \left(\mathbb{E}_{t-1}^{\theta} - \mathbb{E}_{t-1}^{w}\right) \left[\mathcal{L}\left(Z_t, f_t\left(w\right)\right)\left\{M_w\right\} - \mathcal{L}\left(Z_t, f_t\left(\theta\right)\right)\left\{M_{\theta}\right\}\right] \\
&\quad + \left(\mathbb{E}_{t-1}^{\theta} - \mathbb{E}_{t-1}^{w}\right) \left[\mathcal{L}\left(Z_t, f_t\left(w\right)\right)\left\{M_w^c\right\} - \mathcal{L}\left(Z_t, f_t\left(\theta\right)\right)\left\{M_{\theta}^c\right\}\right] \\
&\leq \left(\mathbb{E}_{t-1}^{\theta} - \mathbb{E}_{t-1}^{w}\right) \Delta_t\left(w, \theta\right)\left\{\left|\Delta_t\left(w, \theta\right)\right| \leq M\right\} \\
&\quad + \left[\mathbb{E}_{t-1}^{\theta} \mathcal{L}\left(Z_t, f_t\left(w\right)\right)\left\{M_w^c\right\} + \mathbb{E}_{t-1}^{w} \mathcal{L}\left(Z_t, f_t\left(\theta\right)\right)\left\{M_{\theta}^c\right\}\right] \\
&\quad [\text{by non-negativity of the loss function}] \\
&= \mathrm{I}_t + \mathrm{II}_t.
\end{aligned}
$$

Summing over t, dividing by T, and taking expectation, for $M > 0$,

$$\begin{aligned}
\mathbb{E}^{\theta}\frac{1}{T}\sum_{t=1}^{T}\mathrm{I}_t &= \mathbb{E}^{\theta}\frac{1}{T}\sum_{t=1}^{T}\int_{\mathcal{Z}}\Delta_t(w,\theta)\{|\Delta_t(w,\theta)|\leq M\}[p_\theta(z|\mathcal{F}_{t-1})-p_w(z|\mathcal{F}_{t-1})]\mu(dz)\\
&\leq \mathbb{E}^{\theta}\frac{1}{T}\sum_{t=1}^{T}M\int_{\mathcal{Z}}|p_\theta(z|\mathcal{F}_{t-1})-p_w(z|\mathcal{F}_{t-1})|\mu(dz)\\
&\leq \mathbb{E}^{\theta}\frac{1}{T}\sum_{t=1}^{T}M\sqrt{2D_t(P_\theta\|P_w)}\\
&\quad \text{[by Pinsker's inequality, e.g. Pollard, 2002, eq.13, p.62]}\\
&\leq M\sqrt{2\mathbb{E}^{\theta}\frac{1}{T}\sum_{t=1}^{T}D_t(P_\theta\|P_w)}\\
&\quad \text{[by Jensen's inequality and concavity of the square root function]}\\
&= M\sqrt{2\frac{1}{T}\mathbb{E}^{\theta}D_{1,T}(P_\theta\|P_w)}.
\end{aligned}$$

Using Holder's inequality, for any t,

$$\begin{aligned}
\mathbb{E}^{\theta}\mathrm{II}_t &\leq \left[\mathbb{E}^{\theta}\mathbb{E}_{t-1}^{\theta}\mathcal{L}(Z_t,f_t(w))^r\right]^{1/r}\left[\mathbb{E}^{\theta}\mathbb{E}_{t-1}^{\theta}\{M_w^c\}\right]^{(r-1)/r}\\
&\quad+\left[\mathbb{E}^{\theta}\mathbb{E}_{t-1}^{w}\mathcal{L}(Z_t,f_t(\theta))^r\right]^{1/r}\left[\mathbb{E}^{\theta}\mathbb{E}_{t-1}^{w}\{M_\theta^c\}\right]^{(r-1)/r}\\
&= o\left(M^{-(r-1)}\right)
\end{aligned}$$

by Condition 2 using the fact that on the r.h.s. the first term in each product is finite while the second term in the product is $o(M^{-r})$ because existence of an r^{th} moment implies tails that are $o(M^{-r})$ (e.g. Serfling, 1980, Lemma 1.14). Hence,

$$\begin{aligned}
\mathbb{E}^{\theta}\frac{1}{T}\sum_{t=1}^{T}(\mathrm{I}_t+\mathrm{II}_t) &\leq M\sqrt{2\frac{1}{T}\mathbb{E}^{\theta}D_{1,T}(P_\theta\|P_w)}+o\left(M^{-(r-1)}\right)\\
&= o\left(\left|\frac{1}{T}\mathbb{E}^{\theta}D_{1,T}(P_\theta\|P_w)\right|^{(r-1)/2r}\right)
\end{aligned}$$

setting $M = o\left(\left|\frac{1}{T}\mathbb{E}^{\theta}D_{1,T}(P_\theta\|P_w)\right|^{-1/2r}\right)$. Taking $\sup_\theta$, and substituting in, an application of Theorem 1 gives the universality result. ■

Proof. [Theorem 3] By Condition 3,

$$\mathbb{E}^{\theta}\ln\sum_{k\in\mathcal{K}}P_{w_k}\left(Z_1^t\right)m(k) \quad\geq\quad \mathbb{E}^{\theta}\ln P_{w_k}\left(Z_1^t\right)+\ln m(k)$$

and we can then proceed exactly as in the proof of Theorem 1 with the extra error term $-\ln m(k)$. ■

Proof. [Theorem 4] By Lemma 3,

$$\begin{aligned}
-\sum_{s=1}^{S}\sum_{t=T_{s-1}+1}^{T_s}\ln p_w\left(Z_{T_s}|\mathcal{F}_{T_s-1}\right) &\leq -\sum_{s=1}^{S}\ln\left[\int_\Theta p_\theta\left(Z_{T_{s-1}+1}^{T_s}|\mathcal{F}_{T_{s-1}}\right)w\left(d\theta\right)\right]\\
&\quad -\sum_{s=2}^{S}\ln\left(\lambda T_{s-1}^{-\alpha}\right)-\sum_{s=1}^{S}\sum_{t=T_{s-1}+1}^{T_s}\ln\left(1-\lambda t^{-\alpha}\right)\\
&\quad \text{[because there is no update at } t=T_0\text{]}\\
&\leq -\sum_{s=1}^{S}\ln\left[\int_\Theta p_\theta\left(Z_{T_{s-1}+1}^{T_s}|\mathcal{F}_{T_{s-1}}\right)w\left(d\theta\right)\right]\\
&\quad \frac{2\lambda}{\sqrt{1-\lambda^2}}\left(1+\frac{T^{1-\alpha}-1}{1-\alpha}\right)+S\ln\left(1/\lambda\right)+\alpha S\ln T
\end{aligned}$$

by (27) (with $S=1$) and (28) in Lemma 5. By Condition 4, as in the proof of Theorem 1,

$$\begin{aligned}
&\sum_{s=1}^{S}\mathbb{E}^{\theta_1^s}\left\{\ln p_{\theta_s}\left(Z_{T_{s-1}+1}^{T_s}|\mathcal{F}_{t-1}\right)-\ln\left[\int_\Theta p_\theta\left(Z_{T_{s-1}+1}^{T_s}|\mathcal{F}_{T_{s-1}}\right)w\left(d\theta\right)\right]\right\}\\
\leq\ &\sum_{s=1}^{S}\inf_{\delta_s>0}\left[\delta_s-\ln w\left(B_{\mathcal{T}_s}\left(\theta_s,\delta_s\right)\right)\right].
\end{aligned}$$

Hence, this display and the previous one implies the result. ■

The following notation will be used in some of the remaining proofs.

Notation 2 *$w_t'(\bullet) := w'(\bullet|\mathcal{F}_t)$ and similarly for $w(\bullet|\mathcal{F}_t)$, where $w(\bullet) := w_0(\bullet) := w(\bullet|\mathcal{F}_0)$; $w'(\bullet|\mathcal{F}_0) =: w'(\bullet) = w(\bullet)$. If u and v are measures such that u is absolutely continuous w.r.t. v, then du/dv stands for the Radon Nikodym derivative of u w.r.t. v.*

Proof. [Theorem 5 and 6] For each $s\in\{1,...,S\}$, define

$$\tilde{u}_{s(v)}\left(d\theta\right)=\tilde{u}_v\left(d\theta\right):=\frac{w\left(d\theta\right)}{w\left(B_v\left(\tilde{\theta}_v,\delta_1^S\right)\right)}I\left\{\theta\in B_v\left(\tilde{\theta}_v,\delta_1^S\right)\right\} \tag{19}$$

where $B_v\left(\tilde{\theta}_v, \delta_1^S\right)$ is as in Condition 5. For any $u_s \in \{\tilde{u}_1, ..., \tilde{u}_V\}$

$$\begin{aligned}
& \mathbb{E}^{\theta_1^S}\sum_{s=1}^{S}\sum_{t\in\mathcal{T}_s}\left[\ln p_{\theta_s}\left(Z_t|\mathcal{F}_{t-1}\right) - \ln p_w\left(Z_t|\mathcal{F}_{t-1}\right)\right] \\
= \; & \mathbb{E}^{\theta_1^S}\sum_{s=1}^{S}\sum_{t\in\mathcal{T}_s}\ln\left[\frac{p_{\theta_s}\left(Z_t|\mathcal{F}_{t-1}\right)}{p_{\theta}\left(Z_t|\mathcal{F}_{t-1}\right)}\right]u_s\left(d\theta\right) \\
& +\mathbb{E}^{\theta_1^S}\sum_{s=1}^{S}\sum_{t\in\mathcal{T}_s}\ln\left[\frac{p_{\theta}\left(Z_t|\mathcal{F}_{t-1}\right)}{p_w\left(Z_t|\mathcal{F}_{t-1}\right)}\right]u_s\left(d\theta\right) \\
\leq \; & \sum_{s=1}^{S}\delta_s + \mathbb{E}^{\theta_1^S}\sum_{s=1}^{S}\sum_{t\in\mathcal{T}_s}\ln\left[\frac{p_{\theta}\left(Z_t|\mathcal{F}_{t-1}\right)}{p_w\left(Z_t|\mathcal{F}_{t-1}\right)}\right]u_s\left(d\theta\right) \qquad (20)
\end{aligned}$$

by Definition of $B_v\left(\tilde{\theta}_v, \delta_1^S\right)$. By (9) and (10), u_s is absolutely continuous w.r.t. w_t' because $\lambda_t\left(0\right) > 0$. Therefore, we can apply Lemma 4,

$$\begin{aligned}
& \mathbb{E}^{\theta_1^S}\sum_{s=1}^{S}\sum_{t\in\mathcal{T}_s}\int_{\Theta}\ln\left(\frac{p_{\theta}\left(Z_t|\mathcal{F}_{t-1}\right)}{\int_{\Theta}p_{\theta'}\left(Z_t|\mathcal{F}_{t-1}\right)w\left(d\theta'|\mathcal{F}_{t-1}\right)}\right)u_s\left(d\theta\right) \\
\leq \; & \sum_{s=1}^{S}\left[\int_{\Theta}\ln\left(\frac{du_s}{dw'_{T_{s-1}-r_s}}\right)du_s - \int_{\Theta}\ln\left(\frac{du_s}{dw'_{T_s}}\right)du_s\right] \qquad (21) \\
& -\sum_{t=1}^{T_1-1}\ln\lambda_t\left(t\right) - \sum_{s=2}^{S}\sum_{t=T_{s-1}+1}^{T_s-1}\ln\lambda_t\left(t\right) - \ln\lambda_T\left(T\right) - \sum_{s=2}^{S}\ln\lambda_{T_{s-1}}\left(T_{s-1}-r_s\right).
\end{aligned}$$

Though the sum for s runs from 1 to S, there are only V different shifts, i.e. $u_s \in \{\tilde{u}_1, ..., \tilde{u}_V\}$. For each s we can choose r_s so that the sum in the brackets in (21) telescopes except for the first and last term of each sequence of shifts of the same kind. Hence, denoting by $\left[T_{v(s)-1}+1, T_{v(s)}\right]$ the

s^{th} time segment such that $u_s = \tilde{u}_v$,

$$\begin{aligned}
& \sum_{s=1}^{S}\left[\int_{\Theta}\ln\left(\frac{du_s}{dw'_{T_{s-1}-r_s}}\right)du_s - \int_{\Theta}\ln\left(\frac{du_s}{dw'_{T_s}}\right)du_s\right] \\
= & \sum_{v=1}^{V}\sum_{s=1}^{S(v)}\left[\int_{\Theta}\ln\left(\frac{d\tilde{u}_v}{dw'_{T_{v(s)-1}-r_{v(s)}}}\right)d\tilde{u}_v - \int_{\Theta}\ln\left(\frac{d\tilde{u}_v}{dw'_{T_{v(s)}}}\right)d\tilde{u}_v\right] \qquad (22) \\
\leq & \sum_{v=1}^{V}\left[\int_{\Theta}\ln\left(\frac{d\tilde{u}_v}{dw'_0}\right)d\tilde{u}_v - \int_{\Theta}\ln\left(\frac{d\tilde{u}_v}{dw'_{T_{S(v)}}}\right)d\tilde{u}_v\right] \\
& [\text{setting } r_{v(s+1)} = T_{v(s+1)-1} - T_{v(s)} \text{ and } r_{v(1)} = T_{v(1)-1} \\
& \text{so that the sum telescopes}] \\
\leq & \sum_{v=1}^{V}\int_{\Theta}\ln\left(\frac{d\tilde{u}_v}{dw'_0}\right)d\tilde{u}_v \\
& [\text{because the second integral in the brackets is positive}] \\
= & -\sum_{v=1}^{V}\ln w\left(B_v\left(\tilde{\theta}_v, \delta_1^S\right)\right)
\end{aligned}$$

substituting (19) and evaluating the integral. To prove the theorems, it is sufficient to bound

$$-\sum_{t=1}^{T_1-1}\ln\lambda_t(t) - \sum_{s=2}^{S}\sum_{t=T_{s-1}+1}^{T_s-1}\ln\lambda_t(t) - \sum_{s=2}^{S}\ln\lambda_{T_{s-1}}(T_{s-1}-r_s) \qquad (23)$$

uniformly in r_s. To this end, for both updates

$$\begin{aligned}
-\sum_{t=1}^{T_1-1}\ln\lambda_t(t) - \sum_{s=2}^{S}\sum_{t=T_{s-1}+1}^{T_s-1}\ln\lambda_t(t) \quad \leq & \quad \sum_{t=S}^{T}\ln\lambda_t(t) \\
& \quad [\text{because } -\ln\lambda_t(t) \text{ is increasing in } t] \\
\leq & \quad \frac{2\lambda}{\sqrt{1-\lambda^2 S^{-2\alpha}}}\left(S^{-\alpha} + \frac{T^{1-\alpha} - S^{1-\alpha}}{1-\alpha}\right)
\end{aligned}$$

by Lemma 5. Now consider

$$\mathrm{I} := \sum_{s=2}^{S}\ln\lambda_{T_{s-1}}(T_{s-1}-r_s)$$

for each update separately. For Theorem 5,

$$\begin{aligned}
\mathrm{I} \quad = & \quad \sum_{s=2}^{S}\ln\left(T_{s-1}^{(1+\alpha)}/\lambda\right) \\
\leq & \quad (S-1)\ln(1/\lambda) + (1+\alpha)(S-1)\ln T
\end{aligned}$$

by (28) in Lemma 5. For Theorem 6, note that

$$\begin{aligned}-\ln\lambda_{T_{s-1}}\left(T_{s-1}-r_s\right) &= \ln\left(1/\lambda\right)+\alpha\ln T_{s-1}+\ln A_{T_{s-1}}+2\ln\left(1+r_s\right)\\ &= \mathrm{I}_s+\mathrm{II}_s+\mathrm{III}_s+\mathrm{IV}_s\end{aligned}$$

and we shall bound the sum of the above, term by term, uniformly in r_s. Trivially,

$$\sum_{s=2}^{S}\mathrm{I}_s=\left(S-1\right)\ln\left(1/\lambda\right).$$

By (28) in Lemma 5

$$\sum_{s=2}^{S}\mathrm{II}_s \leq \alpha\left(S-1\right)\ln T.$$

By (29) in Lemma 5,

$$\sum_{s=2}^{S}\mathrm{III}_s\leq 0.$$

Finally,

$$\begin{aligned}\sum_{s=1}^{S}\mathrm{IV}_s &= 2\sum_{s=2}^{S}\ln\left(1+r_s\right)\\ &\leq 2\left(S-1\right)\ln\left(1+\frac{1}{S-1}\sum_{s=2}^{S}r_s\right)\\ &\quad\text{[by concavity and Jensen's inequality]}\\ &= 2\left(S-1\right)\ln\left(1+\frac{1}{S-1}\sum_{v=1}^{V}\sum_{s=1}^{S(v)}r_{v(s)}\right)\end{aligned}$$

by the same arguments and notation in (22). Recalling that in (22) we set $r_{v(s+1)}=T_{v(s+1)-1}-T_{v(s)}$ and $r_{v(1)}=T_{v(1)-1}$, we bound

$$\begin{aligned}\sum_{s=1}^{S(v)}r_{v(s)} &= T_{v(1)-1}+\sum_{s=2}^{S(v)}\left(T_{v(s)-1}-T_{v(s-1)}\right)\\ &= T_{v(S(v))-1}+\sum_{s=1}^{S(v)}\left(T_{v(s)-1}-T_{v(s)}\right)\\ &\leq \left(T-1\right)-S\left(v\right)\end{aligned}$$

where we have bounded $T_{v(S(v))-1}\leq\left(T-1\right)$ and $\left(T_{v(s)-1}-T_{v(s)}\right)\leq -1$ because each segment $\left[T_{v(s)-1},T_{v(s)}\right]$ must have length at least one. Summing over v and substituting in the previous display,

$$\begin{aligned}\sum_{s=1}^{S}\mathrm{IV}_s &\leq 2\left(S-1\right)\ln\left(1+\sum_{v=1}^{V}\frac{\left(T-1\right)-S\left(v\right)}{S-1}\right)\\ &\leq 2\left(S-1\right)\ln\left(\frac{V\left(T-1\right)}{S-1}\right)\end{aligned}$$

because $\sum_{v=1}^{V}S\left(v\right)/\left(S-1\right)>1$. Putting everything together gives the bound for I under Theorem 6. The results are then given backing up all the previous bounds and substituting them in (23), substituting this equation and (22) in (21) and finally substituting (21) in (20). ■

Proof. [Theorem 7] Define $\Delta_t\left(w,P\right):=\mathcal{L}\left(Z_t,f_t\left(w\right)\right)-\mathcal{L}\left(Z_t,f_t\left(P\right)\right)$ and $M_P:=\left\{\mathcal{L}\left(Z_t,f_t\left(P\right)\right)\leq M\right\}$ and M_P^c for its complement. Then, following the proof of Theorem 2, using Condition 6 instead of Condition 2, and the just defined notation,

$$\begin{aligned}\mathbb{E}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{t-1}\Delta_t\left(w,P\right) &= \mathbb{E}\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}_{t-1}-\mathbb{E}_{t-1}^{w}\right)\Delta_t\left(w,P\right)\\ &\quad+\mathbb{E}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{t-1}^{w}\Delta_t\left(w,P\right)\\ &\leq M\sqrt{2\mathbb{E}D_{1,T}\left(P\|P_w\right)/T}\\ &\quad+\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}_{t-1}\mathcal{L}\left(Z_t,f_t\left(w\right)\right)\left\{M_w^c\right\}+\mathbb{E}_{t-1}^{w}\mathcal{L}\left(Z_t,f_t\left(\theta\right)\right)\left\{M_P^c\right\}\right]\\ &= \mathrm{I}+\mathrm{II}.\end{aligned}$$

To bound I, by the properties of the KL divergence

$$\begin{aligned}\mathbb{E}D_{1,T}\left(P\|P_w\right) &= \mathbb{E}D_{1,T}\left(P\|P_\theta\right)+\mathbb{E}\sum_{t=1}^{T}\mathbb{E}_{t-1}\ln\frac{p_\theta\left(Z_t|\mathcal{F}_{t-1}\right)}{p_w\left(Z_t|\mathcal{F}_{t-1}\right)}\\ &= \mathbb{E}D_{1,T}\left(P\|P_\theta\right)+\mathbb{E}\left[\ln p_\theta\left(Z_1^T\right)-p_w\left(Z_1^T\right)\right]\\ &\leq \mathbb{E}D_{1,T}\left(P\|P_\theta\right)+\delta-\ln w\left(B_T\left(\theta\right)\right) \qquad (24)\end{aligned}$$

by (4). To bound II, mutatis mutandis, as in the proof of Theorem 2, by Condition 6,

$$\begin{aligned}\mathbb{E}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{t-1}\Delta_t\left(w,P\right) &\leq M\sqrt{2\mathbb{E}D_{1,T}\left(P\|P_w\right)/T}+o\left(M^{-\left(r-1\right)}\right)\\ &= o\left(\left|\mathbb{E}D_{1,T}\left(P\|P_w\right)/T\right|^{\left(r-1\right)/2r}\right)\end{aligned}$$

setting $M=o\left(\left|\mathbb{E}D_{1,T}\left(P\|P_w\right)/T\right|^{-1/2r}\right)$. Substituting (24) inside and taking $\inf_\theta\inf_\delta$ gives the result. ■

A.1 Technical Lemmata

Lemma 2 *For any $T \in \mathbb{N}$, for the predictor p_w defined by (2) and (3),*

$$p_w\left(Z_T|\mathcal{F}_{T-1}\right) = \frac{\int_\Theta p_\theta\left(Z_1^T\right) w\left(d\theta\right)}{\prod_{t=1}^{T-1} p_w\left(Z_t|\mathcal{F}_{t-1}\right)}$$

implying

$$p_w\left(Z_T|\mathcal{F}_{T-1}\right) = \frac{\int_\Theta p_\theta\left(Z_1^T\right) w\left(d\theta\right)}{\int_\Theta p_\theta\left(Z_1^{T-1}\right) w\left(d\theta\right)}.$$

Proof. [Lemma 2] Note that (3) can be written as

$$w\left(d\theta|\mathcal{F}_T\right) = \frac{w\left(d\theta|\mathcal{F}_{T-1}\right) p_\theta\left(Z_T|\mathcal{F}_{T-1}\right)}{p_w\left(Z_T|\mathcal{F}_{T-1}\right)}$$

so that

$$\begin{aligned} p_w\left(Z_T|\mathcal{F}_{T-1}\right) &= \int_\Theta p_\theta\left(Z_T|\mathcal{F}_{T-1}\right) w\left(d\theta|\mathcal{F}_{T-1}\right) \\ &= \frac{\int_\Theta p_\theta\left(Z_{T-1}^T|\mathcal{F}_{T-2}\right) w\left(d\theta|\mathcal{F}_{T-2}\right)}{p_w\left(Z_{T-1}|\mathcal{F}_{T-2}\right)} \end{aligned}$$

and the first equality follows by recursion. Finally,

$$\begin{aligned} p_w\left(Z_T|\mathcal{F}_{T-1}\right) &= \frac{\int_\Theta p_\theta\left(Z_1^T\right) w\left(d\theta\right)}{p_w\left(Z_{T-1}|\mathcal{F}_{t-2}\right) \prod_{t=1}^{T-2} p_w\left(Z_t|\mathcal{F}_{t-1}\right)} \\ & \quad \text{[factoring out } p_w\left(Z_{T-1}|\mathcal{F}_{t-2}\right)\text{]} \\ &= \frac{\prod_{t=1}^{T-2} p_w\left(Z_t|\mathcal{F}_{t-1}\right)}{\int_\Theta p_\theta\left(Z_1^{T-1}\right) w\left(d\theta\right)} \frac{\int_\Theta p_\theta\left(Z_1^T\right) w\left(d\theta\right)}{\prod_{t=1}^{T-2} p_w\left(Z_t|\mathcal{F}_{t-1}\right)} \end{aligned}$$

substituting the first inequality of the lemma. The result then follows by obvious cancellation of terms. ■

Lemma 3 *For any $t \in \mathbb{N}$, suppose*

$$w\left(d\theta|\mathcal{F}_t\right) = \left(1-\lambda_t\right) w'\left(d\theta|\mathcal{F}_t\right) + \lambda_t w\left(d\theta\right) \tag{25}$$

where $\lambda_t \in (0,1)$ and $w'\left(d\theta|\mathcal{F}_t\right)$ is as in (10). Then,

$$\begin{aligned} -\sum_{t=T_{s-1}+1}^{T_s} \ln p_w\left(Z_{T_s}|\mathcal{F}_{T_s-1}\right) &\leq -\ln \int_\Theta p_\theta\left(Z_{T_{s-1}+1}^{T_s}|\mathcal{F}_{T_{s-1}}\right) w\left(d\theta\right) \\ & \quad -\ln \lambda_{T_{s-1}} - \sum_{t=T_{s-1}+1}^{T_s} \ln\left(1-\lambda_t\right) \end{aligned}$$

Proof. [Lemma 3] By (25)

$$\begin{aligned}
p_w\left(Z_{T_s}|\mathcal{F}_{T_s-1}\right) &= \int_\Theta p_\theta\left(Z_{T_s}|\mathcal{F}_{T_s-1}\right)\left[\left(1-\lambda_{T_s-1}\right)w'\left(d\theta|\mathcal{F}_{T_s-1}\right)+\lambda_{T_s-1}w\left(d\theta\right)\right] \\
&\geq \left(1-\lambda_{T_s-1}\right)\int_\Theta p_\theta\left(Z_{T_s}|\mathcal{F}_{T_s-1}\right)w'\left(d\theta|\mathcal{F}_{T_s-1}\right) \\
&\quad \text{[by positivity of each single term]} \\
&= \left(1-\lambda_{T_s-1}\right)\int_\Theta \frac{p_\theta\left(Z_{T_s}|\mathcal{F}_{T_s-1}\right)p_\theta\left(Z_{T_s-1}|\mathcal{F}_{T_s-2}\right)w\left(d\theta|\mathcal{F}_{T_s-2}\right)}{p_w\left(Z_{T_s-1}|\mathcal{F}_{T_s-2}\right)} \\
&\quad \text{[by (10)]} \\
&\geq \lambda_{T_{s-1}}\prod_{t=T_{s-1}+1}^{T_s}\left(1-\lambda_t\right)\int_\Theta \frac{p_\theta\left(Z_{T_{s-1}+1}^{T_s}|\mathcal{F}_{T_{s-1}}\right)w\left(d\theta\right)}{\prod_{t=T_{s-1}+1}^{T_s-1}p_w\left(Z_t|\mathcal{F}_{t-1}\right)}
\end{aligned}$$

iterating and lower bounding $w'\left(d\theta|\mathcal{F}_{T_{s-1}}\right)$ with $\lambda_{T_{s-1}}w\left(d\theta\right)$. Taking $-\ln$ on both sides,

$$\begin{aligned}
-\ln p_w\left(Z_{T_s}|\mathcal{F}_{T_s-1}\right) &\leq -\ln\int_\Theta p_\theta\left(Z_{T_{s-1}+1}^{T_s}|\mathcal{F}_{T_{s-1}}\right)w\left(d\theta\right)+\sum_{t=T_{s-1}+1}^{T_s-1}\ln p_w\left(Z_t|\mathcal{F}_{t-1}\right) \\
&\quad -\ln\lambda_{T_{s-1}}-\sum_{t=T_{s-1}+1}^{T_s}\ln\left(1-\lambda_t\right),
\end{aligned}$$

and rearranging

$$\begin{aligned}
-\sum_{t=T_{s-1}+1}^{T_s}\ln p_w\left(Z_{T_s}|\mathcal{F}_{T_s-1}\right) &\leq -\ln\int_\Theta p_\theta\left(Z_{T_{s-1}+1}^{T_s}|\mathcal{F}_{T_{s-1}}\right)w\left(d\theta\right) \\
&\quad -\ln\lambda_{T_{s-1}}-\sum_{t=T_{s-1}+1}^{T_s}\ln\left(1-\lambda_t\right).
\end{aligned}$$

■

Lemma 4 *For $s=1,...,S$, suppose u_s is a measure on Θ, absolutely continuous w.r.t. $w\left(\bullet|\mathcal{F}_{t-1}\right)$, $t\in\mathcal{T}_s$. Then for $r\geq 0$, and $s>1$,*

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_s}\int_\Theta \ln\left(\frac{p_\theta\left(Z_t|\mathcal{F}_{t-1}\right)}{\int_\Theta p_{\theta'}\left(Z_t|\mathcal{F}_{t-1}\right)w\left(d\theta'|\mathcal{F}_{t-1}\right)}\right)u_s\left(d\theta\right) \\
\leq\; &\int_\Theta \ln\left(\frac{du_s}{dw'_{T_{s-1}-r}}\right)du_s-\int_\Theta \ln\left(\frac{du_s}{dw'_{T_s}}\right)du_s \\
&-\sum_{t=T_{s-1}+1}^{T_s-1}\ln\lambda_t\left(t\right)-\ln\lambda_{T_{s-1}}\left(T_{s-1}-r\right).
\end{aligned}$$

and for $s = 1$

$$\begin{aligned} & \sum_{t=1}^{T_1} \int_\Theta \ln \left(\frac{p_\theta (Z_t|\mathcal{F}_{t-1})}{\int_\Theta p_{\theta'} (Z_t|\mathcal{F}_{t-1}) \, dw (\theta'|\mathcal{F}_{t-1})} \right) u_1 (d\theta) \\ \leq & \int_\Theta \ln \left(\frac{du_1 (\theta)}{dw'_{T_{s-1}-r}} \right) du_1 - \int_\Theta \ln \left(\frac{du_s}{dw'_{T_1}} \right) du_1 \\ & - \sum_{t=1}^{T_1 - 1} \ln \lambda_t (t) . \end{aligned}$$

Proof. [Lemma 4] By (10) and the Radon Nikodym Theorem,

$$\begin{aligned} \mathrm{I}_t (s) \quad := & \int_\Theta \ln \left(\frac{p_\theta (Z_t|\mathcal{F}_{t-1})}{\int_\Theta p_{\theta'} (Z_t|\mathcal{F}_{t-1}) \, dw (\theta'|\mathcal{F}_{t-1})} \right) u_s (d\theta) \\ = & \int_\Theta \ln \left(\frac{dw'_t}{dw_{t-1}} \right) du_s \\ \leq & \int_\Theta \ln \left(\frac{dw'_t}{\lambda_{t-1} (t-1-r) \, dw'_{t-1-r}} \right) u_s (d\theta) \end{aligned} \tag{26}$$

by (9) noting that all the terms in the summation in (9) are positive. Writing $\ln \lambda_{t-1-r} (t-1-r)$ outside and summing over t, with $r = 0$ when $T_{s-1} + 1 < t \leq T_s$ and leaving r arbitrary but fixed when $t = T_{s-1} + 1$ and $s > 1$,

$$\begin{aligned} \sum_{t \in \mathcal{T}_s} \mathrm{I}_t (s) \quad \leq & \int_\Theta \ln \left(\frac{dw'_{T_s}}{dw'_{T_{s-1}-r}} \right) du_s - \sum_{t=T_{s-1}+2}^{T_s} \ln \lambda_{t-1} (t-1) - \ln \lambda_{T_{s-1}} (T_{s-1} - r) \\ = & \int_\Theta \ln \left(\frac{du_s}{dw'_{T_{s-1}-r}} \right) du_s - \int_\Theta \ln \left(\frac{du_s}{dw'_{T_s}} \right) du_s \\ & - \sum_{t=T_{s-1}+2}^{T_s} \ln \lambda_{t-1} (t-1) - \ln \lambda_{T_{s-1}} (T_{s-1} - r) . \end{aligned}$$

We still need to deal with the case $t = 1$. In this case, note that $w_0 = w'_0$ so that we can directly substitute in (26) without incurring the extra error $-\ln \lambda_0 (0)$ at the first trial (note that a fortiori, $r = 0$). By a change of variable in the sums, the results follow. ■

Lemma 5 *Using the notation of Theorem 4, for* $\alpha \geq 0$ *and* $\lambda \in (0, 1)$,

$$\sum_{t=S}^{T} \ln \left(1 - \lambda t^{-\alpha} \right) < \frac{2\lambda}{\sqrt{1 - \lambda^2 S^{-2\alpha}}} \left(S^{-\alpha} + \frac{T^{1-\alpha} - S^{1-\alpha}}{1 - \alpha} \right) \tag{27}$$

$$-\sum_{s=2}^{S} \ln \left(\lambda T_{s-1}^{-\alpha} \right) \leq (S-1) \ln (1/\lambda) + \alpha (S-1) \ln T \tag{28}$$

$$\sum_{s=2}^{S} \ln A_{T_{s-1}} \leq 0 \tag{29}$$

Proof. [Lemma 5] For $x \in [0,1]$, Taylor expansion of $\ln(1-\lambda x)$ around $x=0$ shows that

$$\begin{aligned} -\ln(1-\lambda x) &= \sum_{i=1}^{\infty} (\lambda x)^i / i \\ &\leq \sqrt{\sum_{i=1}^{\infty} (\lambda x)^{2i} \sum_{i=1}^{\infty} i^{-2}} \\ &= \sqrt{\frac{(\lambda x)^2}{1-(\lambda x)^2} \frac{\pi^2}{6}} \\ &< \frac{2\lambda x}{\sqrt{1-(\lambda x)^2}}. \end{aligned} \tag{30}$$

Hence,

$$\begin{aligned} -\sum_{t=S}^{T} \ln\left(1-\lambda t^{-\alpha}\right) &< \frac{2\lambda}{\sqrt{1-\lambda^2 S^{-2\alpha}}} \sum_{t=S}^{T} t^{-\alpha} \\ &\quad \text{[by (30)]} \\ &= \frac{2\lambda}{\sqrt{1-\lambda^2 S^{-2\alpha}}} \left(S^{-\alpha} + \sum_{t=S+1}^{T} t^{-\alpha}\right) \\ &\leq \frac{2\lambda}{\sqrt{1-\lambda^2 S^{-2\alpha}}} \left(S^{-\alpha} + \int_{S}^{T} t^{-\alpha} dt\right) \\ &= \frac{2\lambda}{\sqrt{1-\lambda^2 S^{-2\alpha}}} \left(S^{-\alpha} + \frac{T^{1-\alpha} - S^{1-\alpha}}{1-\alpha}\right) \end{aligned}$$

by a simple integral bound for the sum, showing (27). The second inequality trivially follows noting that $T > T_{S-1}$. To show (29), note that

$$\begin{aligned} \sum_{r=0}^{t-1} (1+t-r)^{-2} &= \sum_{r=2}^{t+1} r^{-2} \\ &\leq \int_{1}^{t+1} r^{-2} dr \\ &= 1-(t+1)^{-1} \end{aligned}$$

using the integral bound for the sum of a decreasing function. Hence,

$$\begin{aligned}\sum_{s=2}^{S}\ln A_{T_{s-1}} &= \sum_{s=2}^{S}\ln\left(\sum_{r=0}^{T_{s-1}-1}(1+T_{s-1}-r)^{-2}\right)\\ &\leq \sum_{s=2}^{S}\ln\left(1-(T_{s-1}+1)^{-1}\right)\\ &\leq 0\end{aligned}$$

because the argument of ln is less than one. ■

References

[1] An, S. and F. Schorfheide (2007) Bayesian Analysis of DSGE Models. Econometric Reviews, 113-172.

[2] Barron A.R. (1988) The Exponential Convergence of Posterior Probabilities with Implications for Bayes Estimators of Density Functions. Department of Statistics Technical Report 7, University of Illinois, Champaign, Illinois. Available from URL: http://www.stat.yale.edu/~arb4/Publications.htm

[3] Barron A.R. (1998) Information-Theoretic Characterization of Bayes Performance and the Choice of Priors in Parametric and Nonparametric Problems. In J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds), Bayesian Statistics 6, 27-52. Oxford University Press.

[4] Barron, A., J. Rissanen and B. Yu (1998) The Minimum Description Length Principle in Coding and Modeling. IEEE Transactions on Information Theory 44, 2743-2760.

[5] Barron, A., M.J. Schervish, and L. Wasserman (1999) The Consistency of Posterior Distributions in Nonparametric Problems. Annals of Statistics 27, 536-561.

[6] Bousquet, O. and M.K. Warmuth (2002) Tracking a Small Set of Experts by Mixing Past Posteriors. Journal of Machine Learning Research 3, 363-396.

[7] Canova, F. and M. Ciccarelli (2004) Forecasting and Turning Point Predictions in a Bayesian Panel VAR Model. Journal of Econometrics 120, 327-359.

[8] Cesa-Bianchi, N. and G. Lugosi (2006) Prediction, Learning, and Games. Cambridge: Cambridge University Press.

[9] Chib, S. (2004) Markov Chain Monte Carlo Technology. In J.E. Gentle, W. Härdle and Y. Mori (eds.) Handbook of Computational Statistics, 71-102. Berlin: Springer.

[10] Chib, S. and F. Nardari and N. Shephard (2006) Analysis of High Dimensional Multivariate Stochastic Volatility Models. Journal of Econometrics 134, 341-371.

[11] Clarke, B. (2007) Information Optimality and Bayesian Modelling. Journal of Econometrics 138, 405-429.

[12] Clarke B. and A.R. Barron (1990) Information Theoretic Asymptotics of Bayes Methods. IEEE Transactions on Information Theory 38, 453-471.

[13] Dawid, A.P. (1984) Statistical Theory. The Prequential Approach. Journal of the Royal Statistical Society, Ser.A 147, 278-292.

[14] Dawid, A.P. (1986) Probability Forecasting. In S. Kotz, N.L. Johnson and C.B. Read (eds.), Encyclopedia of Statistical Sciences Vol. 7, 210-218. Wiley.

[15] Diaconis, P. and D. Freedman (1986) On the Consistency of Bayes Estimates. Annals of Statistics 14, 1-67.

[16] Evans, M. T. Swartz (1995) Methods for Approximating Integrals in Statistics with Special Emphasis on Bayesian Integration Problems. Statistical Science 10, 254-272.

[17] Geweke, J. (1989) Bayesian Inference in Econometric Models Using Monte Carlo Integration. Econometrica 57, 1317-1339.

[18] Geweke, J. (2005) Contemporary Bayesian Econometrics and Statistics. Hoboken, NJ: Wiley.

[19] Gourieroux, G., A. Monfort and A. Trognon (1984) Pseudo Maximum Likelihood Methods: Theory. Econometrica 52, 681-700.

[20] Hamilton J.D. (2005) Regime-Switching Models. In S.N. Durlauf and L.E. Blume (eds.) The New Palgrave Dictionary of Economics, forthcoming.

[21] Haussler, D. (1997) A general Minimax Result for Relative Entropy. IEEE Transactions on Information Theory 43, 1276-1280.

[22] Haussler, D. and M. Opper (1997) Mutual Information, Metric Entropy and Cumulative Relative Entropy Risk. Annals of Statistics 25, 2451-2492.

[23] Herbster, M. and M.K. Warmuth (1998) Tracking the Best Expert. Machine Learning 32, 151-178.

[24] Hutter, M. (2005) Universal Artificial Intelligence. Berlin: Springer.

[25] Kass R.E. and A.E. Raftery (1995) Bayes Factors. Journal of the American Statistical Association 90, 773-795.

[26] Madigan D. and A.E. Raftery (1994) Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. Journal of the American Statistical Association 89, 1535-1546.

[27] Merhav, N. and M. Feder (1998) Universal Prediction. IEEE Transactions on Information Theory 44, 2124-2147.

[28] Pesaran, M.H. and A. Timmermann (2005) Real-Time Econometrics. Econometric Theory 21, 212-231.

[29] Pesaran, M.H., D. Pettenuzzo, and A. Timmermann (2006) Forecasting Time Series Subject to Multiple Structural Breaks. Review of Economic Studies 73, 1057-1084.

[30] Phillips, P.C.B. and W. Ploberger (1996) An Asymptotic Theory of Bayesian Inference for Time Series. Econometrica 64, 381-412.

[31] Pollard, D. (2002) A User's Guide to Measure Theoretic Probability. Cambridge: Cambridge University Press.

[32] Rissanen, J (1986) Stochastic Complexity and Modeling. Annals of Statistics 14, 1080-1100.

[33] Sancetta, A. (2007) Online Forecast Combinations of Distributions: Worst Case Bounds. Journal of Econometrics, in press.

[34] Schorfheide, F. (2007) Bayesian Methods in Macroeconometrics. In S.N. Durlauf and L.E. Blume (eds.) The New Palgrave Dictionary of Economics, forthcoming.

[35] Serfling, R.J. (1980) Approximation Theorems of Mathematical Statistics. New York: Wiley.

[36] Sims, C.A. and T. Zha (1998) Bayesian Methods for Dynamic Multivariate Models. International Economic Review 39, 949-968.

[37] Strasser, H. (1981) Consistency of Maximum Likelihood and Bayes Estimates. Annals of Statistics 9, 1107-1113.

[38] Timmermann, A. (2006) Forecast Combinations. In G. Elliott, C.W.J. Granger and A. Timmermann, Handbook of Economic Forecasting. Amsterdam: North-Holland.

[39] Yang, Y. (2004) Combining Forecasting Procedures: Some Theoretical Results. Econometric Theory 20, 176-222.

[40] Zellner, A. (1988) Optimal Information Processing and Bayes's Theorem. Journal of the American Statistical Association 42, 278-284.

[41] Zellner, A. (2002) Information Processing and Bayesian Analysis. Information and Entropy Econometrics. Journal of Econometrics 107, 41-50.