Bayesian Analysis (2012) **7**, Number 1, pp. 1–36

Universality of Bayesian Predictions

Alessio Sancetta*

Abstract. 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 misspecification.

Key Words: Bayesian methods, loss function, model averaging, structural change, universal prediction.

1 Introduction

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), Zellner (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 requires only a mild condition on the prior, i.e. the prior needs to be information dense at the "true value" (e.g. Barron (1988), Barron (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 in this paper are related. However, the present discussion will also bring together ideas and results from a rich literature on information theory (e.g. Merhav and Feder (1998)), artificial intelligence (e.g. Cesa-Bianchi and Lugosi (2006), 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 result presented here will be followed by a discussion of related references.

The focus of the paper is theoretical. However, its conclusions have clear practical implications 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 parame-

*The author was lecturer at the University of Cambridge until 2008. He has then worked on several algorithmic trading positions in the City of London. He now works as a freelancer. http://sites.google.com/site/wwwsancetta/

 DOI:10.1214/12-BA701

ter, the optimal choice of prior is known to be related to the information matrix (i.e. an exponential tilt of Jeffreys prior) and more details can be given (see Clarke and Barron (1990), for exact conditions), but will not be discussed here.

1.1 Background and Notation

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. Denote by $P_\theta(\bullet|\mathcal{F}_{t-1})$ the law of Z_t conditional on the sigma algebra $\mathcal{F}_{t-1}$ generated by $(Z_s)_{s<t}$, where $\mathcal{F}_0$ 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 $z_1^t := (z_1, ..., z_t)$ and the above are understood as distribution functions. 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 integrated likelihood is

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

where w is a prior probability measure on subsets of Θ. Then, the Bayesian predictive density at z_t is

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

where $0/0 := 0$. In a prediction context, the sequential loss incurred by using $p_w\left(z_t|\mathcal{F}_{t-1}\right)$ in place of $p_\theta\left(z_t|\mathcal{F}_{t-1}\right)$ can be measured by the relative entropy (or interchangeably 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. $P_\theta(\bullet|\mathcal{F}_{t-1})$ 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 prequential KL divergence (e.g. Dawid (1984), Dawid (1992), Dawid (1998)). Previously, for ease of notation, z is used in place of z_t. Letting $\mathbb{E}^\theta$ be the unconditional expectation w.r.t. P_θ, it follows that

$$\mathbb{E}^\theta D_{1,T}\left(P_\theta\|P_w\right) = \int_{\mathcal{Z}^T} p_\theta\left(z_1^T\right) \ln\left(\frac{p_\theta\left(z_1^T\right)}{p_w\left(z_1^T\right)}\right) \mu\left(dz_1^T\right),$$

which is the joint KL divergence. The Bayesian estimator $p_w\left(z_1^T\right)$ arises as the solution to the problem $\inf_P \int_\Theta \mathbb{E}^\theta D_{1,T}\left(P_\theta\|P\right) w\left(d\theta\right)$ where the inf runs over all distribution functions.

Here, universality of prediction shall be defined as follows:

Definition 1. *Predictions based on p_w are universal with respect to $\{P_\theta : \theta \in \Theta\}$ if*

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

Definition 1 is quite forgiving in the sense that it only requires the KL divergence to be sublinear in T, i.e. $o\left(T\right)$. Hence, the term universality is understood in a wide sense in order to accommodate subsequent results to be presented in due course. It is well known that for the KL divergence the exact rate of growth is $O\left(\ln T\right)$ for most regular cases (see Clarke and Barron (1994), for details).

The δ-information neighbourhood of P_θ is denoted by

$$B_T\left(\theta,\delta\right):=\left\{\theta'\in\Theta:\mathbb{E}^\theta D_{1,T}\left(P_\theta\|P_{\theta'}\right)\leq\delta\right\}, \tag{2}$$

or for notational convenience $B_T\left(\theta\right)=B_T\left(\theta,\delta\right)$, whichever is felt to be more appropriate for the situation. The prior w is said to be information dense (at θ) if it assigns strictly positive probability to each information neighbourhood of size δ, i.e. $w\left(B_T\left(\theta,\delta\right)\right)>0$ for any $\delta>0$. 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 $B_T\left(\theta,\delta\right)$ is in terms of either the individual or the joint relative entropy divided by T (e.g. Barron (1998)). For reasons that will become apparent later, we are working with the joint entropy. Hence, information balls of joint entropy less than or equal to δ in this paper would correspond to balls of entropy less than δ/T in the literature.

Information denseness of w is related to the following quantity:

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

where $R_T\left(\theta\right)/T$ is the 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 $T\to\infty$. It can be shown that if w is information dense, then, $R_T\left(\theta\right)/T\to 0$ as $T\to\infty$ (Lemma 2). The following condition ensures that $\sup_{\theta\in\Theta}\mathbb{E}^\theta D_{1,T}\left(P_\theta\|P_w\right)$ goes to zero.

Condition 1.

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

Hence, we can summarize the above remarks through the following well-known result (e.g. Barron (1998)).

Theorem 1. *Using the notation in (2)*

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

so that under Condition 1, predictions based on p_w are universal, i.e.

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

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 $o(T)$ as $T \to \infty$.

As a simple application of Theorem 1, 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 the need to worry about the possible presence of a unit root, and Theorem 1 gives a finite sample upper bound for the loss in the prediction. For example, in the Holder continuity case to be discussed in (29) (e.g. X_t is Gaussian noise, Cauchy, etc.), the resolvability index is $O(\ln T/T)$.

The proof is just a consequence of the chain rule property of the KL divergence. It is instructive to sketch the proof of Theorem 1, as it will be needed later.

Proof.[Theorem 1] By definition,

$$p_w(z_t|\mathcal{F}_{t-1}) = \int_\Theta p_\theta(z_t|\mathcal{F}_{t-1})\, w(d\theta|\mathcal{F}_{t-1}) \tag{3}$$

where

$$w(d\theta|\mathcal{F}_t) = \frac{w(d\theta|\mathcal{F}_{t-1})\, p_\theta(Z_t|\mathcal{F}_{t-1})}{\int_\Theta w(d\theta|\mathcal{F}_{t-1})\, p_\theta(Z_t|\mathcal{F}_{t-1})} \tag{4}$$

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

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

The above display together with (3) imply that

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

so

$$\sum_{t=1}^{T} \ln p_w(Z_t|\mathcal{F}_{t-1}) = \ln p_w\left(Z_1^T\right), \tag{5}$$

because the sum telescopes and $\mathcal{F}_0$ is trivial. Choosing a ball $B(\theta) := B_T(\theta)$ as in (2), one can bound the expectation of the above display,

$$\begin{aligned}
\mathbb{E}^\theta \ln \int_\Theta p_{\theta'}\left(Z_1^T\right) w(d\theta') &\geq \mathbb{E}^\theta \ln \int_{B(\theta)} p_{\theta'}\left(Z_1^T\right) w(d\theta') \\
&\quad \text{[because } p_\theta\left(Z_1^T\right) \text{ is non-negative]} \\
&\geq \mathbb{E}^\theta \ln\left(p_\theta\left(Z_1^t\right)\right) - \delta + \ln w(B(\theta))
\end{aligned} \tag{6}$$

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,

$$\begin{aligned} \mathbb{E}^{\theta} D_{1,T}\left(P_{\theta} \| P_w\right) &= \mathbb{E}^{\theta} \sum_{t=1}^{T} \mathbb{E}_{t-1}^{\theta}\left[\ln\left(p_{\theta}\left(Z_t|\mathcal{F}_{t-1}\right)\right) - \ln\left(p_w\left(Z_t|\mathcal{F}_{t-1}\right)\right)\right] \\ &= \mathbb{E}^{\theta}\left[\ln p_{\theta}\left(Z_1^T\right) - \ln p_w\left(Z_1^T\right)\right] \\ &\quad \text{[by (5)]} \\ &\leq \delta - \ln w\left(B\left(\theta\right)\right) \end{aligned}$$

by (6). 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. ■

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

$$\mathbb{E}^{\theta} \ln p_{\theta}\left(Z_1^T\right) \leq \mathbb{E}^{\theta} \ln\left(\int_{A_T} p_{\theta'}\left(Z_1^T\right) \frac{w\left(d\theta'\right)}{w\left(A_T\right)}\right) + \delta_T T \tag{7}$$

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

The following summarizes the above remarks:

Lemma 2. *An information dense prior w (at θ) implies $\lim_{T\to\infty} R_T\left(\theta\right)/T = 0$ and the latter implies (7) with $\lim_{T\to\infty}\left\{\delta_T T - \ln w\left(A_T\right)\right\}/T = 0$.*

In practice, the 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 the default condition. Moreover, for two results to be stated (Theorem 6 and 7), (7) 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 (2), the infimum $R_T\left(\theta\right)$ is obtained by $\delta = 0$ if Θ is finite and w puts strictly positive mass to each element of Θ (see the proof of Theorem 4, for details). Section 6.6 provides remarks on how to check Condition 1 in an important special case.

The plan of the paper is as follows. Section 2 illustrates an important consequence of universality for prediction under loss functions satisfying a moment bound. Sections 3 and 4 show universality results for model averaging and predictions in a non-stationary environment. Section 5 shows how the results of the paper change under model misspecification. Further discussion including remarks about the conditions can be found

in Section 6 . While outlines of proofs are included in the text, proofs of technical lemmata are relegated to the appendix.

2 Predictions for Arbitrary Loss Functions

The KL divergence satisfies a chain rule for jointly distributed random variables and is minimized (i.e. equal to zero) only when its arguments are the same. These are the essential properties used in the proof of Theorem 1. Another crucial property is that the KL divergence provides an upper bound for the L_1 norm of two densities. Let p and q be densities absolutely continuous with respect to a dominating measure μ. By Pinsker's inequality, the L_1 norm is bounded as follows,

$$\left(\int |p-q|\, pd\mu\right)^2 \leq 2\int \ln\left(\frac{p}{q}\right) pd\mu \tag{8}$$

(Pinsker (1964), Csiszar (1967)). Using the bound (8) we can show that universality has important implications for convergence of Bayesian predictions based on loss functions satisfying some moment conditions.

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). The infeasible 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

$$f_{t+1}(\theta) := \arg\inf_{f\in\mathfrak{F}} \mathbb{E}_t^\theta \mathcal{L}(Z_{t+1}, f). \tag{9}$$

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 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 1. *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,*

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

as the conditional mean is the minimizer of the conditional mean square error (e.g. Harvey (1993)).

Since θ is unknown, in (9) the feasible prediction is obtained by replacing 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 prediction

$$f_{t+1}(w) := \arg\inf_{f\in\mathfrak{F}} \mathbb{E}_t^w \mathcal{L}(Z_{t+1}, f) \tag{10}$$

where $\mathbb{E}_t^w$ stands for expectation with respect to $P_w(\bullet|\mathcal{F}_t)$. We shall see that, as a consequence of universality, this prediction satisfies some desirable properties. Note that $\mathbb{E}_{t-1}^\theta[\mathcal{L}(Z_t, f_t) - \mathcal{L}(Z_t, f_t(\theta))] \geq 0$ by construction, because $f_t(\theta)$ is the predictor that minimizes the loss $\mathcal{L}$ under expectation w.r.t. $P_\theta(\bullet|\mathcal{F}_{t-1})$. In particular, using (8), the goal is to show that universality implies

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

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

Condition 2. *There is an $r > 1$ such that*

$$\sup_{\theta\in\Theta}\sup_{t>0} \mathbb{E}^\theta\left[\mathbb{E}_{t-1}^\theta \mathcal{L}(Z_t, f_t(w))^r + \mathbb{E}_{t-1}^w \mathcal{L}(Z_t, f_t(\theta))^r\right] < \infty.$$

If Θ is compact and the loss function uniformly integrable in t with respect to $(P_\theta)_{\theta\in\Theta}$, then we only need to worry about establishing a moment bound for all θ. Further remarks on Condition 2 will be found in Section 6.7.

We have the following result:

Theorem 3. *Under Condition 2,*

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

Hence, if Condition 1 holds as well, the r.h.s. converges to zero.

Remark 1. *Theorem 3 says that (10) leads to an average conditional prediction error asymptotically equal (in $L_1(P_\theta)$) to the average conditional prediction error obtained using the unfeasible predictions $f_1(\theta), ..., f_T(\theta)$. It is possible to write a proper upper bound in terms of constants that depend on the moments of the loss function only.*

Merhav and Feder (1998) show how to relate the left hand side of Theorem 3 to the joint relative entropy in the case of bounded loss functions. (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.

The proof for bounded loss uses the fact that the loss is non-negative and that $f_t(w)$ is the minimizer of $\mathbb{E}_{t-1}^{w}\mathcal{L}(Z_t, f)$. Hence, adding and subtracting

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

one can bound $\mathbb{E}_{t-1}^{\theta}\left[\mathcal{L}(Z_t, f_t(w)) - \mathcal{L}(Z_t, f_t(\theta))\right]$ with the L_1 distance between p_θ and p_w (conditional on $\mathcal{F}_{t-1}$). By (8), this can be bounded by the square root of the relative entropy. Hence, we can invoke Theorem 1. The slower convergence (dependent on an r moment of the loss function) results from truncating an unbounded loss. Here are the details:

Proof. [Theorem 3] Define $\Delta_t(w,\theta) := \mathcal{L}(Z_t, f_t(w)) - \mathcal{L}(Z_t, f_t(\theta))$.
Then $\mathbb{E}_{t-1}^{w}\Delta_t(w,\theta) \leq 0$ because $f_t(w)$ is the minimizer of $\mathbb{E}_{t-1}^{w}\mathcal{L}(Z_t, f)$. Define the sets

$$M_w := \{\mathcal{L}(Z_t, f_t(w)) \leq M\}$$

and

$$M_\theta := \{\mathcal{L}(Z_t, f_t(\theta)) \leq M\}$$

and denote their complements by M_w^c and M_θ^c. If A is a set, directly use A in place of its indicator function. By this remark, adding and subtracting $\mathbb{E}_{t-1}^{w}\Delta_t(w,\theta)$,

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

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

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

Taking expectation inside the sum, bounding the empirical average with the supremum and using Holder's inequality,

$$\begin{aligned}
\sup_{\theta\in\Theta}\mathbb{E}^{\theta}\frac{1}{T}\sum_{t=1}^{T}\mathrm{II}_t &\leq \sup_{\theta\in\Theta}\sup_{t>0}\left[\mathbb{E}^{\theta}\mathbb{E}^{\theta}_{t-1}\mathcal{L}(Z_t,f_t(w))^r\right]^{1/r}\left[\mathbb{E}^{\theta}\mathbb{E}^{\theta}_{t-1}\{M_w^c\}\right]^{(r-1)/r} \\
&\quad +\sup_{\theta\in\Theta}\sup_{t>0}\left[\mathbb{E}^{\theta}\mathbb{E}^{w}_{t-1}\mathcal{L}(Z_t,f_t(\theta))^r\right]^{1/r}\left[\mathbb{E}^{\theta}\mathbb{E}^{w}_{t-1}\{M_{\theta}^c\}\right]^{(r-1)/r} \\
&= o\left(M^{-(r-1)}\right)
\end{aligned}$$

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 the 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. ■

The rest of the paper will focus on extending Theorem 1 to different scenarios. However, as a consequence of the established universality of predictions, results for arbitrary loss will be stated as corollaries without proof.

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 $\mathcal{K} := \{1, , 2..., K\}$. The Bayesian predictive density under model uncertainty is given by

$$p_m\left(Z_t|\mathcal{F}_{t-1}\right) := \sum_{k\in\mathcal{K}} p_{w_k}\left(Z_t|\mathcal{F}_{t-1}\right) m\left(k|\mathcal{F}_{t-1}\right) \tag{11}$$

where

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

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

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

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

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

Condition 3. *$m\left(k\right) > 0$ for any $k \in \mathcal{K}$.*

Hence, we can state the following:

Theorem 4. *We have the following upper bound,*

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

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

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

Proof. [Theorem 4] By Condition 3,

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

as each term in the sum is positive. We can then proceed exactly as in the proof of Theorem 1 with the extra error term $-\ln m\left(k\right)$. ■

Corollary 1. *Let $\mathbb{E}^m$ be expectation with respect to P_m and $f_{t+1}(m)$ as in (10) using $\mathbb{E}^m$. If*

$$\max_{k\in\mathcal{K}}\sup_{\theta\in\Theta_k}\sup_{t>0}\left[\mathbb{E}_{t-1}^{\theta}\mathcal{L}\left(Z_t,f_t\left(m\right)\right)^r+\mathbb{E}_{t-1}^{m}\mathcal{L}\left(Z_t,f_t\left(\theta\right)\right)^r\right]<\infty$$

for some $r>1$, then,

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

Moreover, since $\mathcal{L}$ is positive, $f_{t+1}(m)=\sum_{k\in\mathcal{K}}m\left(k|\mathcal{F}_{t-1}\right)f_{t+1}\left(w_k\right)$, where $f_{t+1}\left(w_k\right)$ is as in (10) using $\mathbb{E}^{w_k}$.

The computational overhead of the above Bayesian model averaging prediction grows linearly in the number of models K.

The stated version of the upper bound is related to results derived in the machine learning and information theory literature (e.g. Vovk (1998), 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.

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 one where there are switches in regimes but we try not to make any assumptions on the dynamics (see Hamilton (2008), for a review of parametric regime switching models). In this case, learning by Bayes' rule needs to involve learning over changing parameters. We are interested in the joint distribution

$$P_{\theta_1^S}\left(Z_1^t\right)=\prod_{s=1}^{S}\prod_{t=T_{s-1}+1}^{T_s}P_{\theta_s}\left(Z_t|\mathcal{F}_{t-1}\right)\tag{12}$$

where $\theta_1^S:=(\theta_1,...,\theta_S)$, and $0=T_0<T_1<...<T_s=T$ are arbitrary, but fixed. For example, the underlying process could be an inhomogeneous Markov chain.

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 2. *Predictions based on p_w are universal for $\left\{P_{\theta_1^S}:\theta_1^S\in\Theta^S\right\}$ over $S\leq T$*

partitions $\mathcal{T}_1, ..., \mathcal{T}_S$ *if*

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

as $T \to \infty$. *(*Θ^S *is the* S *Cartesian product of* Θ*.)*

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 a priori, we need to introduce a prior on the probability of changes. We 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 $\{0, 1, 2, ..., t\}$, so that $\sum_{r=0}^{t} \lambda_t(t-r) = 1$. Then, we mix past posteriors using $\lambda_t(r)$ as mixing density:

$$w\left(d\theta|\mathcal{F}_t\right) = \sum_{r=0}^{t} \lambda_t(t-r)\, w'\left(d\theta|\mathcal{F}_{t-r}\right) \tag{13}$$

where

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

and

$$w'\left(d\theta|\mathcal{F}_t\right) = \frac{p_\theta\left(Z_t|\mathcal{F}_{t-1}\right) w\left(d\theta|\mathcal{F}_{t-1}\right)}{\int_\Theta p_\theta\left(Z_t|\mathcal{F}_{t-1}\right) w\left(d\theta|\mathcal{F}_{t-1}\right)}. \tag{14}$$

This updating algorithm has been studied by Bousquet and Warmuth (2002) for countable Θ. The update has a clear Bayesian interpretation: with probability $\lambda_t(r)$ the posterior of θ at time t is equal to the posterior $dw'\left(\theta|\mathcal{F}_r\right)$ 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. Data within each regime are generated by the same θ. When $r = 0$ we are taken back to the prior, which corresponds to the start of a new regime that has not previously occurred. Hence, the main idea is to keep some positive probability on past posteriors that contain possibly relevant information for the future. This intuition will be formalized in the proofs.

Note that the use of a prior on the regimes further complicates any computational issues. In fact, at each t, the computational burden scales linearly with the support of $\lambda_t(r)$, i.e. the cardinality of $\{r \in \mathbb{N} : \lambda_t(r) > 0\}$.

We shall use $D_{\mathcal{T}_s}\left(P_\theta \| P_{\theta'}\right) := D_{T_{s-1}+1, T_s}\left(P_\theta \| P_{\theta'}\right)$ for the prequential KL divergence 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*

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

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 based on Bousquet and Warmuth (2002) will be considered. The first case makes no assumption on the type of changes, and only assumes that there are $S-1$ changes. 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. The second case assumes that there are $S-1$ shifts in the parameter, but 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.

4.1 Shrinking towards the Prior

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 (13) simplifies to

$$w(d\theta|\mathcal{F}_t)=\left(1-\lambda t^{-\alpha}\right)w'(d\theta|\mathcal{F}_t)+\lambda t^{-\alpha}w(d\theta). \tag{15}$$

Theorem 5. *Using (15), for any segments $\mathcal{T}_1,...,\mathcal{T}_S$,*

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

so that predictions based on P_w are universal under Condition 4 if $S\ln T=o(T)$ and $\alpha>0$.

Remark 2. *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 5 is monotonically decreasing in α. Increasing α does however increase the last term in the bound, i.e. $\alpha S\ln T$.*

Proof. [Theorem 5] The main intuition in the proof is that we can break down the negative log-likelihood of the Bayesian prediction into S blocks over arbitrary time subsets, as long as we keep some positive weight on the related posterior update. Lemma

9 in the appendix, formalizes this idea. Hence,

$$\begin{aligned}
-\sum_{s=1}^{S}\sum_{t=T_{s-1}+1}^{T_s}\ln p_w\left(Z_{T_s}|\mathcal{F}_{T_s-1}\right) \quad &\leq \quad -\sum_{s=1}^{S}\ln\left[\int_\Theta p_\theta\left(Z_{T_{s-1}+1}^{T_s}|\mathcal{F}_{T_{s-1}}\right)w\left(d\theta\right)\right] \\
&\quad -\sum_{s=2}^{S}\ln\left(\lambda T_{s-1}^{-\alpha}\right)-\sum_{s=1}^{S}\sum_{t=T_{s-1}+1}^{T_s}\ln\left(1-\lambda t^{-\alpha}\right) \\
&\quad \text{[by Lemma 9,} \\
&\quad \text{and because there is no update at } t=T_0\text{]} \\
&\leq \quad -\sum_{s=1}^{S}\ln\left[\int_\Theta p_\theta\left(Z_{T_{s-1}+1}^{T_s}|\mathcal{F}_{T_{s-1}}\right)w\left(d\theta\right)\right] \\
&\quad +\frac{2\lambda}{\sqrt{1-\lambda^2}}\left(1+\frac{T^{1-\alpha}-1}{1-\alpha}\right)+S\ln\left(1/\lambda\right) \\
&\quad +\alpha S\ln T
\end{aligned}$$

by (34) (with $S=1$) and (35) in Lemma 12. By Condition 4, as in the proof of Theorem 1,

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

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

In the bound of Theorem 5, α and λ are free parameters whose choice can be based on prior knowledge or subjective beliefs. If S is of large order, we could minimize the bound setting λ close to one and α close to zero. This suggests that as the number of shifts relative to T increases, 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. see 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 5,

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

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 in the hope of reducing the resolvability index. This issue will be addressed next.

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

Consider now the case of a 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 within V regimes. In particular, to set up 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". Intuitively, past information should be helpful, and we may improve Theorem 5 letting $\lambda_t(r) > 0$ for any $r \leq t$. This is 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,

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

Remark 3. *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 (29).

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:

$$w(d\theta|\mathcal{F}_t) = \left(1 - \lambda t^{-\alpha}\right) w'(d\theta|\mathcal{F}_t) + \sum_{r=1}^{t} \frac{\lambda t^{-\alpha}}{t} w'(d\theta|\mathcal{F}_{t-r}). \tag{16}$$

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

Theorem 6. *Using (16), 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* $V \leq S$,

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

where

$$error\left(T, S, \alpha, \lambda\right) := \frac{2\lambda}{\sqrt{1-\lambda^2 S^{-2\alpha}}} \left(S^{-\alpha} + \frac{T^{1-\alpha} - S^{1-\alpha}}{1-\alpha}\right) + S \ln\left(1/\lambda\right) + \left(1+\alpha\right) S \ln T,$$

so that the prediction is universal under Condition 5 if $S \ln T = o\left(T\right)$ *and* $\alpha > 0$.

Remark 4. *Theorem 6 leads to a decrease in the resolvability index when* V *is fixed and* $S \to \infty$. *Comparing with Theorem 5, this comes at the extra cost of an error term* $S \ln T$, *but with an improvement in*

$$\frac{2\lambda}{\sqrt{1-\lambda^2 S^{-2\alpha}}} \left(S^{-\alpha} + \frac{T^{1-\alpha} - S^{1-\alpha}}{1-\alpha}\right). \tag{17}$$

Section 6.8 provides further remarks on the improvement of the resolvability index, in a special case. When Θ is finite, Bousquet and Warmuth (2002) provide encouraging simulation evidence in favor of mixing past posteriors using $\lambda_t\left(r\right) > 0$ $\left(r \in [0, t]\right)$ 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 5). According to these remarks, the mixing update in (16) 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 (15).

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

$$w\left(d\theta | \mathcal{F}_t\right) = \left(1 - \lambda t^{-\alpha}\right) w'\left(d\theta | \mathcal{F}_t\right) + \sum_{r=1}^{t} \frac{\lambda t^{-\alpha}}{A_t \left(1+r\right)^2} w'\left(d\theta | \mathcal{F}_{t-r}\right). \tag{18}$$

Theorem 7. *Using (18) instead of (16), in Theorem 6, we have*

$$\begin{aligned} error\left(T, S, \alpha, \lambda\right): \quad = \quad & \frac{2\lambda}{\sqrt{1-\lambda^2 S^{-2\alpha}}} \left(S^{-\alpha} + \frac{T^{1-\alpha} - S^{1-\alpha}}{1-\alpha}\right) \\ & + S \ln\left(1/\lambda\right) + \alpha S \ln T + 2S \ln\left(\frac{V\left(T-1\right)}{S-1}\right), \end{aligned}$$

so that the prediction is universal under Condition 5 if $S \ln T = o\left(T\right)$ *and* $\alpha > 0$.

Remark 5. *Theorem 7 shows that the extra cost $S \ln T$ in Theorem 6 can be reduced to $2S \ln\left(\frac{V(T-1)}{S-1}\right)$ if we use (18) instead of (16).*

A slight modification of Condition 2 can be used to deal with predictions under general loss functions.

Corollary 2. *Suppose that there is an $r > 1$, such that for any partition $\{\mathcal{T}_s : s > 0\}$,*

$$\sup_{\theta_1^S \in \Theta^S} \sup_{s \leq S} \sup_{t \in \mathcal{T}_s} \mathbb{E}^{\theta_1^S} \left[\mathbb{E}_{t-1}^{\theta_s} \mathcal{L}\left(Z_t, f_t\left(w\right)\right)^r + \mathbb{E}_{t-1}^{w} \mathcal{L}\left(Z_t, f_t\left(\theta_s\right)\right)^r \right] < \infty.$$

Here $\mathbb{E}_{t-1}^{w}$ is understood as expectation with respect to the predictive density based on mixtures of past posteriors, and similarly for $f_t\left(w\right)$. Then, universality as in Definition 2 implies

$$\sup_{\theta_1^S \in \Theta^S} \frac{1}{T} \mathbb{E}^{\theta_1^S} \sum_{s=1}^{S} \sum_{t=T_{s-1}+1}^{T_s} \mathbb{E}_{t-1}^{\theta_s} \left[\mathcal{L}\left(Z_t, f_t\left(w\right)\right) - \mathcal{L}\left(Z_t, f_t\left(\theta_s\right)\right) \right] \to 0.$$

Proof. [Theorem 6 and 7] Introduce the following notation: $w_t'\left(\bullet\right) := w'\left(\bullet|\mathcal{F}_t\right)$ and similarly for $w\left(\bullet|\mathcal{F}_t\right)$, where $w\left(\bullet\right) := w_0\left(\bullet\right) := w\left(\bullet|\mathcal{F}_0\right)$; $w'\left(\bullet|\mathcal{F}_0\right) =: w'\left(\bullet\right) = w\left(\bullet\right)$. 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.

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 $u_s \in \{\tilde{u}_1, ..., \tilde{u}_V\}$, adding and subtracting $\int_\Theta \ln p_\theta\left(Z_t|\mathcal{F}_{t-1}\right) u_s\left(d\theta\right)$,

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

by Definition of $B_v\left(\tilde{\theta}_v, \delta_1^S\right)$. By (13) and (14), u_s is absolutely continuous w.r.t. w_t' because $\lambda_t\left(0\right) > 0$. Therefore, we can apply Lemma 10:

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

Although 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 $u_s = \tilde{u}_v$, and, with abuse of notation, letting T_v be the last time $\tilde{\theta}_v$ was the true parameter,

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

substituting (19) and evaluating the integral. Once we insert (22) in (21), to prove the theorems, it is sufficient to bound

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

with the constraints used above, i.e. $r_{v(s+1)} = T_{v(s+1)-1} - T_{v(s)}$ and $r_{v(1)} = T_{v(1)-1}$. These quantities are bounded by Lemma 11 in the appendix, by elementary inequalities and some algebra. Once again, the trick is to make sure that we assign positive weight to the information contained in each posterior. ■

Mutatis mutandis, Theorem 5, 6 and 7 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 restrict attention to Θ being finite.

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 face only estimation error. However, when expectation is taken with respect to a probability $P \notin \{P_\theta : \theta \in \Theta\}$, we will also incur an approximation error, hence universality might not be achieved. The approximation error can be characterized in terms of the relative entropy. With no loss of generality, 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. Use $\mathbb{E}$ for (unconditional) expectation w.r.t. P. We shall review the previous results in the light of estimation error.

Condition 6. *Define*

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

Then,

$$\sup_{t>0} \mathbb{E}\left[\mathbb{E}_{t-1} \mathcal{L}\left(Z_t, f_t\left(w\right)\right)^r + \mathbb{E}_{t-1}^w \mathcal{L}\left(Z_t, f_t\left(P\right)\right)^r\right] < \infty$$

for some $r > 1$.

Then, we have an extra error term due to the approximation error.

Theorem 8.

$$\mathbb{E} D_t\left(P\|P_\theta\right) \leq \frac{\inf_{\theta\in\Theta} \inf_\delta \left\{\mathbb{E} D_{1,T}\left(P\|P_\theta\right) + \delta - \ln w\left(B_T\left(\theta, \delta\right)\right)\right\}}{T},$$

and, under Condition 6,

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

Remark 6. *By the following inequality*

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

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. $\inf_{\theta\in\Theta}\mathbb{E}D_{1,T}\left(P\|P_\theta\right)/T$.

Proof. [Theorem 8] The proof follows along the same lines as the proof of Theorem 3 with P_θ replaced by P. After having used Pinsker's inequality to bound the total variation between P and P_w it is enough to note that

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

by (2). The final details are left to the reader. ■

6 Discussion

6.1 Implications of Universality

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

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

(Shannon (1948)). By the properties of the KL divergence, the mutual information is minimized w.r.t. Q by P_w, i.e.

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

for any Q (Aitchison (1975)). Hence, the minimizer of the mutual information is the Bayes risk (e.g. Clarke and Barron (1994), 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 predictive distribution P_w (Haussler (1997), Theorem 1). Hence, 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 $\mathcal{Z}^T$. Given that $D_{1,T}(P_\theta \| P_w) \geq 0$, universality implies $L_1(P_\theta)$ convergence of the prequential KL distance, which in turn implies its convergence in P_θ-probability for any $\theta \in \Theta$.

6.2 Averaging Estimators as Solutions to Bayes Risk

The Bayesian predictive density can be directly derived as a solution to the Bayes risk for the relative entropy. Divergence functions other than the relative entropy (e.g. Hellinger, Chi-square, etc.) give different forms of averaging estimator where the weight on the likelihood is either inflated (e.g. Chi-square divergence) or deflated (Hellinger divergence). With different aims in mind, Corcuera and Giummol (1999), Zellner (2002), and Clarke and Yuan (2010) explicitly look at these modified Bayesian estimators. In general, the modified predictive density takes the form

$$p_w\left(z_1^T; \eta\right) = \frac{\left(\int_\Theta \left[p_\theta\left(z_1^T\right)\right]^\eta w(d\theta)\right)^\kappa}{C_T}, \tag{25}$$

$$C_T := \int_{\mathcal{Z}^T} \left(\int_\Theta \left[p_\theta\left(z_1^T\right)\right]^\eta w(d\theta)\right)^\kappa \mu\left(dz_1^T\right),$$

where $\eta, \kappa > 0$ and C_T is just a constant of integration. The case $\eta = 1/\kappa = 1/2$ is the solution to the Hellinger divergence, while $\eta = 2$, $\kappa = 1$ is the solution to the Chi-square divergence. It can be shown that (25) is asymptotically optimal for the KL divergence when $\eta\kappa = 1$ and $\kappa \geq 1$. The crucial step is to show that (6) holds. To see this, take logs of (25) and write $\left[p_\theta\left(z_1^T\right)\right]^\eta = \exp\left\{\eta \ln p_\theta\left(z_1^T\right)\right\}$ so that, using the same arguments as

in (6),

$$\begin{aligned}\mathbb{E}^{\theta}\ln p_{w}\left(Z_{1}^{T};\eta\right) &= \kappa\mathbb{E}^{\theta}\ln\int_{\Theta}\exp\left\{\eta\ln p_{\theta'}\left(Z_{1}^{T}\right)\right\}w\left(d\theta'\right)-\ln C_{T}\\ &\geq \kappa\eta\mathbb{E}^{\theta}\ln p_{\theta}\left(Z_{1}^{T}\right)-\delta+\ln w\left(B_{T}\left(\theta,\kappa\eta\delta\right)\right)\\ &\quad \text{[because } \kappa\eta=1 \text{ and assuming } C_{T}\leq 1\text{, for the moment]}\\ &= \mathbb{E}^{\theta}\ln p_{\theta}\left(Z_{1}^{T}\right)-\delta+\ln w\left(B_{T}\left(\theta,\delta\right)\right).\end{aligned}$$

Note that, for $\kappa \geq 1$, by Jensen's inequality,

$$\begin{aligned}C_{T} &\leq \int_{\mathcal{Z}^{T}}\int_{\Theta}\left[p_{\theta}\left(z_{1}^{T}\right)\right]^{\eta\kappa}w\left(d\theta\right)\mu\left(dz_{1}^{T}\right)\\ &= 1\end{aligned}$$

when $\eta\kappa = 1$. Hence, the solution to other divergence functions (e.g. Hellinger) can be asymptotically optimal for the KL divergence. The other way around does not appear to be true.

It is interesting to note that when $\eta \to \infty$, the posterior puts all its weight at one point, i.e. the maximum likelihood estimator. Hence, the Laplace approximation to the predictive density is simply obtained by replacing the usual Bayesian posterior with the inflated with $\eta \to \infty$, i.e. all weight is given to the evidence provided by the data relative to the prior (see Zellner (2002), for more on this interpretation). There are some relations between the above discussion and the concept of learning rate and realizable prediction usually employed in the machine learning literature. The interested reader it referred to Haussler et al. (1998), and Vovk (1998) (see also Cesa-Bianchi and Lugosi (2006)).

6.3 Relation to Worst-Case Bounds

The results presented here are also related to competitive online statistics (Vovk (2001)), which in the machine learning literature are usually referred to as predictions with expert advice. There, the focus is on worst-case bounds for

$$D_{1,T}^{(obs)}\left(P_{\theta}\|P_{\theta'}\right):=\sum_{t=1}^{T}\left[\ln\left(p_{\theta}\left(Z_{t}|\mathcal{F}_{t-1}\right)\right)-\ln\left(p_{\theta'}\left(Z_{t}|\mathcal{F}_{t-1}\right)\right)\right] \tag{26}$$

for any data sequence $Z_1, ..., Z_T$ and $T > 0$, which we call the observed joint relative entropy. In particular, it is assumed that nature outputs $Z_1, ..., Z_T$ in an adversarial game where the statistician is required to issue a prediction $p_w\left(\bullet|\mathcal{F}_{t-1}\right)$ before nature outputs Z_t. It can be shown that (26) is bounded by $\inf_{\delta>0}\left\{\delta-\ln w\left(B_{T}^{(obs)}\left(\theta,\delta\right)\right)\right\}$ where

$$B_{T}^{(obs)}\left(\theta,\delta\right):=\left\{\theta'\in\Theta:D_{1,T}^{(obs)}\left(P_{\theta}\|P_{\theta'}\right)\leq\delta\right\} \tag{27}$$

(using the arguments in the proof of Lemma 2 and Theorem 1). In the simplest case of prediction with expert advice, Θ is finite, the prior is uniform over Θ, and the upper

bound for (26) over all bounded data sequences simplifies to $\ln K$, where K is the cardinality of Θ (see Theorem 4 for an application; there the finite set is denoted by $\mathcal{K}$ rather than Θ). It is thus obvious that we can turn many of the results in this paper into worst-case results by simply changing the definition of information neighbour in (2) into that in (27). Note that (27) is a random ball (unless we take the sup over all data sequences). Nevertheless, in some cases, we can control its size ex ante, though asymptotically.

Example 2. *Suppose that, for $t > 0$, Z_t is conditionally distributed as Gaussian with mean θZ_{t-1} and variance one. From (26), by simple algebra, deduce that*

$$B_T^{(obs)}(\theta, \delta) = \left\{ \theta' \in \Theta : \sum_{t=1}^{T} Z_{t-1}(\theta - \theta') \left[Z_t - \frac{(\theta' + \theta)}{2} Z_{t-1} \right] \leq \delta \right\},$$

where Z_0 is fixed (recall that $\mathcal{F}_0$ is trivial). Given, $Z_0, ..., Z_{T-1}$ we can solve for the set of θ' satisfying the inequality in $B_T^{(obs)}(\theta, \delta)$. Moreover, by the law of large numbers, for $|\theta| < 1$,

$$\frac{1}{T}\left(1 - \mathbb{E}^{\theta}\right) \sum_{t=1}^{T} Z_{t-1}(\theta - \theta') \left[Z_t - \frac{(\theta' + \theta)}{2} Z_{t-1} \right] \to 0$$

almost surely (e.g. Brockwell and Davis (1991)). Hence, the information neighbourhood (2) still provides some useful information even in the worst-case scenario. Clearly, we are assuming that nature follows a "well behaved stochastic process" to output Z_t.

For Example 2, a worst-case bound gives an error equal to infinity (see Theorem 1 in Vovk (2001)). It seems that this problem can only be overcome by giving up worst-case bounds. Then, it is possible to derive bounds for the conditional mean loss (Vovk (2001), footnote 5, p.34) or the mean loss as shown in Theorem 3.

This can be achieved by replacing the information neighbourhood (2) with a prequential neighbourhood

$$B_T^{(preq)}(\theta, \delta) := \{\theta' \in \Theta : D_{1,T}(P_\theta \| P_{\theta'}) \leq \delta\}$$

which is the information neighbourhood based on the prequential KL divergence. As remarked by Dawid (1998) the optimality property related to this criterion is prequential efficiency. Prequential efficiency requires almost sure convergence of the prequential KL divergence divided by T, while universality (as discussed here) only requires L_1 convergence of the T standardized prequential KL divergence. Hence, the quantities $D_{1,T}^{(obs)}(P_\theta \| P_{\theta'})$ $D_{1,T}(P_\theta \| P_{\theta'})$ and $\mathbb{E}^{\theta} D_{1,T}(P_\theta \| P_{\theta'})$ require, respectively, control over the sets $B_T^{(obs)}(\theta, \delta)$, $B_T^{(preq)}(\theta, \delta)$ and $B_T(\theta, \delta)$, where $B_T(\theta, \delta) \subseteq B_T^{(preq)}(\theta, \delta) \subseteq B_T^{(obs)}(\theta, \delta)$, almost surely. The smaller the set, the smaller the error in the bound.

6.4 Finite Sample Limitations of Bayesian Predictions Over Competing Approaches

This paper shows that Bayesian predictions possess some desirable properties in terms of universality. Other desirable properties like Prequential Efficiency are discussed by Dawid (e.g. Dawid (1984), Dawid (1992)) in the case of predictions based on a mixture of distributions. However, in finite samples, there are examples of procedures that are superior. Shtarkov (1987) provides details of a density prediction that is optimal in finite samples (i.e. achieves the minimax regret). Cesa-Bianchi and Lugosi (2001) provide theoretical evidence that mixture algorithms like Bayesian predictions are not the best possible in finite samples. Clarke (2003), Wong and Clarke (2004) and Clarke and Clarke (2009) provide finite sample empirical evidence in favour of methods that combine both model-based and empirical approaches, especially when the target model is not in the reference class. In general, outside the realm of prediction with loss based on KL divergence, the results in this paper might be optimal only asymptotically (e.g. Clarke (2007)). Hence, the theoretical soundness of arguments based on Bayesian predictions may come at the cost of finite sample loss.

6.5 Prediction over Multiple Steps Ahead

One of the many issues not discussed includes 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 (see 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 the 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

$$\mathbb{E}_{t-1}^{\theta} \ln \frac{p_\theta\left(Z_t^{t+h}|\mathcal{F}_{t-1}\right)}{p_w\left(Z_t^{t+h}|\mathcal{F}_{t-1}\right)} = \mathbb{E}_{t-1}^{\theta} \ln \frac{p_\theta\left(Z_1^{t+h}\right)}{p_w\left(Z_1^{t+h}\right)} - \mathbb{E}_{t-1}^{\theta} \ln \left[\frac{p_\theta\left(Z_1^{t-1}\right)}{p_w\left(Z_1^{t-1}\right)}\right] \{t > 1\} \tag{28}$$

using (1) and similar steps as in the proof of Theorem 1. 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:

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

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, in doing so, (28) 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.

6.6 Remarks on Condition 1

The verification of Condition 1 requires smoothness of the joint 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 $t \in \mathbb{N}$

$$\mathbb{E}^{\theta}\left[\ln p_{\theta'}\left(Z_t|\mathcal{F}_{t-1}\right) - \ln p_{\theta}\left(Z_t|\mathcal{F}_{t-1}\right)\right] \leq b\left|\theta' - \theta\right|^{a} \tag{29}$$

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

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

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 joint relative entropy divided by T converges to zero at the rate $\ln T/T$ for any Holder's continuous class of expected conditional log-likelihoods. In order of magnitude, this recovers the result of Clarke and Barron (1994). Using additional smoothness conditions, these authors derive more detailed results that are then linked to Jeffreys prior.

Note that in (29) we may have $b \asymp t$ (l.h.s. and r.h.s. are of the same order) (as for Example 2 when $\theta = 1$). However, the resolvability index will only be affected by a multiplicative constant. To put (29) into perspective, note that the differentiability of the expected conditional log-likelihood per observation is stronger than (29). The following is a prototypical example where standard maximum likelihood methods are known to fail for some parameter values.

Example 3. *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, (29) holds with $a = 1$, while p_{θ} is not differentiable at $\theta = 0$. In a frequentist context, the Hellinger differentiability is often used to overcome this problem (e.g. Pollard (2003), ch 4).*

6.7 Remarks on Condition 2

Condition 2 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. Choose $\mathfrak{F}$ to contain all the functions 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.

Example 4. *Suppose $p_\theta(Z_t|\mathcal{F}_{t-1}) = p_\theta(Z_t|Z_{t-1})$ is a Markov transition density. Restrict $\mathfrak{F}$ to contain only functions f such that $|f(z)| \leq g(z) = 1 + b|z|^a$ for some $a, b > 0$. Suppose that the loss function can be bounded as follows $\mathcal{L}(z, f) \leq |z| + |f|$, e.g. absolute loss. Then, to check Condition 2 note that*

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

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

6.8 Improvement on the Resolvability Index of Theorem 7 over Theorem 5

Consider the Holder's continuity condition in (29) and the same prior as given in its discussion. For simplicity, suppose that all the time segments $\mathcal{T}_s$ have the 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 5,

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

substituting the minimizer $\delta = a^{-1}$. Clearly, if S is of large order this quantity will be large. On the other hand, in Theorem 7 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.

6.9 Further Remarks

This paper provides a comprehensive set of results for universal prediction using Bayes' rule. The conditions used restrict Θ only implicitly. For Condition 1 to hold, Θ cannot be completely arbitrary, but the restrictions on Θ are quite mild.

The relative improvement on the resolvability index when we mix past posteriors (and not just the prior, i.e. (15)) 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 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, (18) already improved upon (16). Perhaps more definite claims could be made if a different method of proof were used.

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 conditions. There are links between the Bayesian information criterion and other measures of complexity like Rissanen's minimum description length principle (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 (see Hutter (2005), for details). Tight estimates of model complexity are the key for tight and explicit rates of convergence of Bayesian predictions.

Appendix 1: Technical Lemmata

Proof. [Lemma 2] Information denseness implies $-\ln w(B_T(\theta,\delta_T T)) < \infty$ for any $\delta_T > 0$. Hence $\delta_T - T^{-1}\ln w(B_T(\theta,\delta_T T))$ can be made arbitrarily small by choosing $\delta_T \to 0$ at a suitable rate as $T \to \infty$. This implies $R_T(\theta)/T \to 0$. To show the last implication, define

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

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

$$\mathbb{E}^\theta D_{1,T}(P_\theta \| P_{w,A_T}) \le \delta_T T \tag{30}$$

which is (7). Setting $B_T(\theta) := B_T(\theta,\delta_T T)$,

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

by definition of $B_T(\theta)$. The above inequality together with (30) imply that $B_T(\theta, \delta_T T) \subseteq A_T(\theta, \delta_T T)$. ■

Lemma 9. *For any $t \in \mathbb{N}$, suppose*

$$w(d\theta|\mathcal{F}_t) = (1-\lambda_t)\, w'(d\theta|\mathcal{F}_t) + \lambda_t w(d\theta) \tag{31}$$

where $\lambda_t \in (0,1)$ and $w'(d\theta|\mathcal{F}_t)$ is as in (14). Then,

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

Proof. [Lemma 9] By (31)

$$\begin{aligned} p_w(Z_{T_s}|\mathcal{F}_{T_s-1}) &= \int_\Theta p_\theta(Z_{T_s}|\mathcal{F}_{T_s-1}) \left[(1-\lambda_{T_s-1})\, w'(d\theta|\mathcal{F}_{T_s-1}) + \lambda_{T_s-1} w(d\theta)\right] \\ &\geq (1-\lambda_{T_s-1}) \int_\Theta p_\theta(Z_{T_s}|\mathcal{F}_{T_s-1})\, w'(d\theta|\mathcal{F}_{T_s-1}) \\ &\quad \text{[by positivity of each term in the brackets]} \\ &= (1-\lambda_{T_s-1}) \int_\Theta \frac{p_\theta(Z_{T_s}|\mathcal{F}_{T_s-1})\, p_\theta(Z_{T_s-1}|\mathcal{F}_{T_s-2})\, w(d\theta|\mathcal{F}_{T_s-2})}{p_w(Z_{T_s-1}|\mathcal{F}_{T_s-2})} \\ &\quad \text{[by (14)]} \\ &\geq \lambda_{T_{s-1}} \prod_{t=T_{s-1}+1}^{T_s} (1-\lambda_t) \int_\Theta \frac{p_\theta\left(Z_{T_{s-1}+1}^{T_s}|\mathcal{F}_{T_{s-1}}\right) w(d\theta)}{\prod_{t=T_{s-1}+1}^{T_s-1} p_w(Z_t|\mathcal{F}_{t-1})} \end{aligned}$$

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

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

and rearranging gives the result. ■

Lemma 10. *For $s = 1, ..., S$, suppose u_s is a measure on Θ, absolutely continuous w.r.t. $w(\bullet|\mathcal{F}_{t-1})$, $t \in \mathcal{T}_s$. Use the notation defined at the beginning of the proof of Theorem 6*

and 7. 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 $s=1$, with $w'_0=w_0$,

$$\begin{aligned}
&\sum_{t=1}^{T_1}\int_\Theta \ln\left(\frac{p_\theta\left(Z_t|\mathcal{F}_{t-1}\right)}{\int_\Theta p_{\theta'}\left(Z_t|\mathcal{F}_{t-1}\right)dw\left(\theta'|\mathcal{F}_{t-1}\right)}\right)u_1\left(d\theta\right)\\
\leq\ &\int_\Theta \ln\left(\frac{du_1\left(\theta\right)}{dw'_0}\right)du_1-\int_\Theta \ln\left(\frac{du_1}{dw'_{T_1}}\right)du_1\\
&-\sum_{t=1}^{T_1-1}\ln\lambda_t\left(t\right).
\end{aligned}$$

Proof. [Lemma 10] By (14) and the Radon Nikodym Theorem,

$$\begin{aligned}
\mathrm{I}_t\left(s\right)\ :=\ &\int_\Theta \ln\left(\frac{p_\theta\left(Z_t|\mathcal{F}_{t-1}\right)}{\int_\Theta p_{\theta'}\left(Z_t|\mathcal{F}_{t-1}\right)dw\left(\theta'|\mathcal{F}_{t-1}\right)}\right)u_s\left(d\theta\right)\\
=\ &\int_\Theta \ln\left(\frac{dw'_t}{dw_{t-1}}\right)du_s\\
\leq\ &\int_\Theta \ln\left(\frac{dw'_t}{\lambda_{t-1}\left(t-1-r\right)dw'_{t-1-r}}\right)u_s\left(d\theta\right)
\end{aligned}\tag{32}$$

by (13) noting that all the terms in the summation in (13) are positive. Writing $\ln\lambda_{t-1-r}\left(t-1-r\right)$ 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 $s>1$,

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

We still need to deal with the case $t=1$. Since $w_0=w'_0$, we can directly substitute in (32) without incurring the extra error $-\ln\lambda_0\left(0\right)$ at the first trial (a fortiori, $r=0$). By a change of variable in the sums, the result follows. ■

Lemma 11. *For Theorem 6, (21) is bounded above by*

$$\frac{2\lambda}{\sqrt{1-\lambda^2 S^{-2\alpha}}}\left(S^{-\alpha}+\frac{T^{1-\alpha}-S^{1-\alpha}}{1-\alpha}\right)+(S-1)\ln(1/\lambda)+(1+\alpha)(S-1)\ln T$$

and for Theorem 7 by

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

Proof. For both Theorems, using the fact that $\lambda_t(t)>0$,

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

using Lemma 12 (note that $\ln\lambda_t(t)=1-\lambda t^{-\alpha}$). The last term $\sum_{s=2}^{S}\ln\lambda_{T_{s-1}}(T_{s-1}-r_s)$, in (21), is bounded differently for Theorem 6 and 7. For Theorem 6, (35) in Lemma 12 gives the first stated bound. For Theorem 7, note that

$$\begin{aligned}
&-\sum_{s=2}^{S}\ln\lambda_{T_{s-1}}(T_{s-1}-r_s)\\
=\;&\sum_{s=2}^{S}\ln(1/\lambda)+\alpha\sum_{s=2}^{S}\ln T_{s-1}+\sum_{s=2}^{S}\ln A_{T_{s-1}}+2\sum_{s=2}^{S}\ln(1+r_s)\\
\leq\;&(S-1)\ln(1/\lambda)+\alpha(S-1)\ln T+2\sum_{s=2}^{S}\ln(1+r_s),
\end{aligned}$$

by (35) and (36) in Lemma 12. We shall bound the last term under the constraints used in Theorem 7, i.e. $r_{v(s+1)}=T_{v(s+1)-1}-T_{v(s)}$ and $r_{v(1)}=T_{v(1)-1}$ used in (22). To this end,

$$\begin{aligned}
2\sum_{s=2}^{S}\ln(1+r_s) &\leq 2(S-1)\ln\left(1+\frac{1}{S-1}\sum_{s=2}^{S}r_s\right)\\
&\quad [\text{ by concavity and Jensen's inequality}]\\
&= 2(S-1)\ln\left(1+\frac{1}{S-1}\sum_{v=1}^{V}\sum_{s=1}^{S(v)}r_{v(s)}\right) \qquad (33)
\end{aligned}$$

by the same arguments and notation in (22). Under the constraints in $r_{v(s)}$,

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

where we have bounded $T_{v(S(v))-1} \leq (T-1)$ 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 a length of at least one. Summing over v and substituting in (33), we have the following bound for (33),

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

because $\sum_{v=1}^{V} S(v)/(S-1) > 1$. ■

Lemma 12. *Using the notation of Theorem 5, for $\alpha \geq 0$ and $\lambda \in (0,1)$,*

$$\sum_{t=S}^{T} \ln\left(1 - \lambda t^{-\alpha}\right) < \frac{2\lambda}{\sqrt{1-\lambda^2 S^{-2\alpha}}}\left(S^{-\alpha} + \frac{T^{1-\alpha} - S^{1-\alpha}}{1-\alpha}\right) \tag{34}$$

$$-\sum_{s=2}^{S} \ln\left(\lambda T_{s-1}^{-\alpha}\right) \leq (S-1)\ln(1/\lambda) + \alpha(S-1)\ln T \tag{35}$$

$$\sum_{s=2}^{S} \ln A_{T_{s-1}} \leq 0. \tag{36}$$

Proof. [Lemma 12] For $x \in [0,1]$, Taylor expansion of $\ln(1-\lambda x)$ around $x = 0$ shows

that

$$\begin{aligned} -\ln(1-\lambda x) &= \sum_{i=1}^{\infty} (\lambda x)^i / i \\ &\leq \sqrt{\sum_{i=1}^{\infty} (\lambda x)^{2i} \sum_{i=1}^{\infty} i^{-2}} \\ &= \sqrt{\frac{(\lambda x)^2}{1-(\lambda x)^2} \frac{\pi^2}{6}} \\ &< \frac{2\lambda x}{\sqrt{1-(\lambda x)^2}}. \end{aligned} \tag{37}$$

Hence,

$$\begin{aligned} -\sum_{t=S}^{T} \ln\left(1-\lambda t^{-\alpha}\right) &< \frac{2\lambda}{\sqrt{1-\lambda^2 S^{-2\alpha}}} \sum_{t=S}^{T} t^{-\alpha} \\ &\quad \text{[by (37)]} \\ &= \frac{2\lambda}{\sqrt{1-\lambda^2 S^{-2\alpha}}} \left(S^{-\alpha} + \sum_{t=S+1}^{T} t^{-\alpha}\right) \\ &\leq \frac{2\lambda}{\sqrt{1-\lambda^2 S^{-2\alpha}}} \left(S^{-\alpha} + \int_S^T t^{-\alpha} dt\right) \\ &= \frac{2\lambda}{\sqrt{1-\lambda^2 S^{-2\alpha}}} \left(S^{-\alpha} + \frac{T^{1-\alpha} - S^{1-\alpha}}{1-\alpha}\right) \end{aligned}$$

by a simple integral bound for the sum, showing (34). The second inequality trivially follows noting that $T > T_{S-1}$. To show (36), 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

Aitchison, J. (1975). "Goodness of Prediction Fit." *Biometrika*, 62: 547–554. 20

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*.
URL http://www.stat.yale.edu/~arb4/Publications.htm 1

— (1998). "Information-Theoretic Characterization of Bayes Performance and the Choice of Priors in Parametric and Nonparametric Problems." In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 6*, 27–52. Oxford: Oxford University Press. 1, 3

Barron, A. R., Rissanen, J., and Yu, B. (1998). "The Minimum Description Length Principle in Coding and Modeling." *IEEE Transactions on Information Theory*, 44: 2743–2760. 27

Barron, A. R., Schervish, M. J., and Wasserman, L. (1999). "The Consistency of Posterior Distributions in Nonparametric Problems." *Annals of Statistics*, 27: 536–561. 3

Bousquet, O. and Warmuth, M. K. (2002). "Tracking a Small Set of Experts by Mixing Past Posteriors." *Journal of Machine Learning Research*, 3: 363–396. 12, 13, 16, 19

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. New York: Springer. 23

Cesa-Bianchi, N. and Lugosi, G. (2001). "Worst-Case Bounds for the Logarithmic Loss of Predictors." *Machine Learning*, 43(3): 247–264. 24

— (2006). *Prediction, Learning, and Games*. Cambridge: Cambridge University Press. 1, 11, 22

Clarke, B. (2003). "Comparing Bayes and Non-Bayes Model Averaging When Model Approximation Error Cannot Be Ignored." *Journal of Machine Learning Research*, 4: 683–712. 24

— (2007). "Information Optimality and Bayesian Modelling." *Journal of Econometrics*, 138: 405–429. 1, 24

Clarke, B. and Barron, A. (1990). "Information Theoretic Asymptotics of Bayes Methods." *IEEE Transactions on Information Theory*, 38: 453–471. 2

— (1994). "Jeffreys' Prior is Asymptotically Least Favourable Under Entropy Risk." *The Journal of Statistical Planning and Inference*, 41: 37–60. 3, 20, 25

Clarke, B. and Yuan, A. (2010). "Reference Priors for Empirical Likelihoods." In Chen, M.-H., Mueller, P., Sun, D., Ye, K., and Dey, D. K. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis: In honor of James O. Berger*, 56–68. New York: Springer. 21

Clarke, J. and Clarke, B. (2009). "Prequential Analysis of Complex Data with Adaptive Model Reselection." *Statistical Analysis and Data Mining*, 2: 274–290. 24

Corcuera, J. M. and Giummol, F. (1999). "A Generalized Bayes Rule for Prediction." *Scandinavian Journal of Statistics*, 26: 265–279. 21

Csiszar, I. (1967). "Information-Type Measures of Divergence of Probability Distributions and Indirect Observations." *Studia Scientiarum Mathematicarum Hungarica*, 2: 299–318. 6

Dawid, A. P. (1984). "Statistical Theory. The Prequential Approach." *Journal of the Royal Statistical Society, Ser. A*, 147: 278–292. 2, 24

— (1992). "Prequential Analysis, Stochastic Complexity and Bayesian Inference." In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 4*. Oxford: Oxford University Press. 2, 24

— (1998). "Discussion of "Information-Theoretic Characterization of Bayes Performance and the Choice of Priors in Parametric and Nonparametric Problems"." In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 6*, 48. Oxford: Oxford University Press. 2, 23

Diaconis, P. and Freedman, D. (1986). "On the Consistency of Bayes Estimates." *Annals of Statistics*, 14: 1–67. 1

Hamilton, J. D. (2008). "Regime Switching Models." In Durlauf, S. N. and Blume, L. E. (eds.), *The New Palgrave Dictionary of Economics*. Basingstoke: Palgrave Macmillan. 11

Harvey, A. (1993). *Time Series Models*. London: Harvester Wheatsheaf. 6

Haussler, D. (1997). "A general Minimax Result for Relative Entropy." *IEEE Transactions on Information Theory*, 43: 1276–1280. 21

Haussler, D. and Opper, M. (1997). "Mutual Information, Metric Entropy and Cumulative Relative Entropy Risk." *Annals of Statistics*, 25: 2451–2492. 20

Haussler, D. J., Kivinen, J., and Warmuth, M. K. (1998). "Sequential Prediction of Individual Sequences under General Loss Functions." *IEEE Transactions on Information Theory*, 44(5): 1906–1925. 22

Herbster, M. and Warmuth, M. K. (1998). "Tracking the Best Expert." *Machine Learning*, 32: 151–178. 19

Hutter, M. (2005). *Universal Artificial Intelligence*. Berlin: Springer. 1, 7, 24, 27

Merhav, N. and Feder, M. (1998). "Universal Prediction." *IEEE Transactions on Information Theory*, 44: 2124–2147. 1, 7

Pinsker, M. S. (1964). *Information and Information Stability of Random Variables and Processes*. San Francisco: Holden-Day. 6

Pollard, D. (2003). *Asymptopia*. Unpublished manuscript.
URL http://www.stat.yale.edu/~pollard/ 25

Rissanen, J. (1986). "Stochastic Complexity and Modeling." *Annals of Statistics*, 14: 1080–1100. 27

Sancetta, A. (2007). "Online Forecast Combinations of Distributions: Worst-Case Bounds." *Journal of Econometrics*, 141: 621–651. 11

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley. 9

Shannon, C. E. (1948). "A Mathematical Theory of Communication." *The Bell System Technical Journal*, 27: 379–423, 623–656. 20

Shtarkov, Y. (1987). "Universal Sequential Coding of Single Messages." *Translated from: Problems in Information Transmission*, 23: 3–17. 24

Strasser, H. (1981). "Consistency of Maximum Likelihood and Bayes Estimates." *Annals of Statistics*, 9: 1107–1113. 1

Timmermann, A. (2006). "Forecast Combinations." In Elliott, G., Granger, C., and Timmermann, A. (eds.), *Handbook of Economic Forecasting*. Amsterdam: North-Holland. 14

Vovk, V. (1998). "A Game of Prediction with Expert Advice." *Journal of Computer and System Sciences*, 56: 153–173. 11, 22

— (2001). "Competitive On-Line Statistics." *International Statistical Review*, 69: 213–248. 22, 23

Wong, H. and Clarke, B. (2004). "Improvement over Bayes Prediction in Small Samples in the Presence of Model Uncertainty." *Canadian Journal of Statistics*, 32: 269–283. 24

Yang, Y. (2004). "Combining Forecasting Procedures: Some Theoretical Results." *Econometric Theory*, 20: 176–222. 11

Zellner, A. (1988). "Optimal Information Processing and Bayes's Theorem. With comments and a reply by the author." *American Statistician*, 42: 278–284. 1

— (2002). "Information Processing and Bayesian Analysis. Information and Entropy Econometrics." *Journal of Econometrics*, 107: 41–50. 1, 21, 22

Acknowledgments

I thank Oliver Linton, Volodya Vovk, and the late Arnold Zellner for comments and/or for suggesting some useful references. I am greatly indebted to two anonymous referees and an associate editor for remarks that led to considerable improvements both in content and presentation.