The Safe Bayesian:
learning the learning rate via the mixability gap

Peter Grünwald

CWI, Amsterdam and Leiden University, The Netherlands
pdg@cwi.nl

Abstract. Standard Bayesian inference can behave suboptimally if the model is wrong. We present a modification of Bayesian inference which continues to achieve good rates with wrong models. Our method adapts the Bayesian learning rate to the data, picking the rate minimizing the cumulative loss of sequential prediction by posterior randomization. Our results can also be used to adapt the learning rate in a PAC-Bayesian context. The results are based on an extension of an inequality due to T. Zhang and others to dependent random variables.

1 Introduction

Problem 1: Bayes when the Model is Wrong Standard Bayesian inference may fail if the probability model $\mathcal{P}$ under consideration is "wrong yet useful". Grünwald and Langford (2007) (GL from now on) exhibit cases in which the posterior never concentrates, putting substantial weight on many "bad" distributions even in the limit of infinite sample size. As a result, predictions based on the posterior remain suboptimal forever. This problem can be addressed by equipping Bayes with a learning rate η as in (Zhang, 2006a). Standard Bayesian inference corresponds to $\eta = 1$; for small enough η, Bayesian inference will become well-behaved again and its predictions will become optimal in the limit. However, picking η too small may lead to an unnecessarily slow convergence rate. The appropriate choice for η depends on the true distribution, which is unknown, and it is unclear how to estimate it from data: GL show that marginalizing out η (as a Bayesian would prefer) does not solve the problem, and picking the η that maximizes the Bayesian marginal likelihood of the data $Z^n = Z_1, \ldots, Z_n$ does not help either (see also Example 3, Example 4 and Figure 1, this paper's **essential picture**).

Problem 2: PAC-Bayesian Learning Rates In statistical learning theory, one consider models Θ of predictors defined relative to some loss function LOSS, e.g. Θ may be a set of classifiers and LOSS may be the 0/1-loss. In *relative PAC-Bayesian bounds* (Audibert, 2004, Zhang, 2006b, Catoni, 2007) one proves frequentist convergence bounds of randomized predictors which depend on some user-specified "prior" distribution over Θ. The bounds are typically optimized by setting the randomized predictor equal to a pseudo-Bayesian posterior at some optimal learning rate η, which once again depends on the unknown true distribution. Algorithms for estimating η from the data have been proposed for special settings (Audibert, 2004), but so far, a general approach has been lacking.

The Safe Bayesian We address both problems at once by picking the $\hat{\eta}$ that maximizes the "sequentially randomized" Bayesian marginal log-likelihood, which for priors with finite support can be reinterpreted as the $\hat{\eta}$ minimizing the cumulative loss of the $\text{HEDGE}(\eta)$ algorithm (Freund and Schapire, 1997). We then predict by the Cesàro average of the Bayesian posteriors at $\hat{\eta}$. We extend this *safe Bayesian* algorithm to the statistical learning case by defining pseudo-probabilities $p_\theta(y \mid x) \propto e^{-\text{LOSS}(y,\theta(x))}$ in the usual manner.

In our first result, Theorem 1, we show that for all η smaller than some "critical" η_{CRIT}, we can expect a small *mixability gap*, a notion reminiscent of Vovk's (1990, 2001) fundamental concept of *mixability* for individual sequence prediction. In our context a small mixability gap means that the expected cumulative log-loss one obtains by *randomizing* according to the posterior is close to the cumulative log-loss one obtains by *mixing* the posterior. If the posterior concentrates, then the mixability gap is small, and we may think of the $\hat{\eta}$ inferred by our algorithm as estimating the largest rate at which the posterior does concentrate. Our main result, Theorem 2 shows that, broadly speaking, the convergence rates achieved by the safe Bayesian algorithm are optimal for the underlying, unknown true distribution in several settings. Specifically, if the model is correct or convex, we perform essentially as well as standard Bayesian inference, which in this case is among the best methods available. Yet when the model is incorrect, in the setting of Grünwald and Langford (2007), unlike standard Bayes, the safe Bayesian posterior does learn to predict optimally, i.e. as well as the single distribution in the model that predicts best.

In Section 2 we introduce notation, concepts and present the safe Bayesian algorithm. Since the algorithm can be applied in a wide variety of contexts (standard Bayes, statistical learning, Hedge-like) this section is, unfortunately, long. In Section 3 we introduce η_{CRIT}, which allows us to give a second, detailed introduction to the results that are to follow. Section 4 gives our first result, relating randomized ("Gibbs") to standard Bayesian prediction and gives, in Figure 1, a crucial picture. Section 5 gives our main result, Theorem 2, showing that the Safe Bayesian algorithm performs comparably to an algorithm that knows the critical learning rate in advance. In Section 6 we compare our results to Grünwald (2011) who already provided a procedure that adapts to η_{CRIT} in a much more restricted setting. In Section 7 we prove Theorem 1 and 2. The latter is built upon Theorem 3, an extension of a PAC-Bayesian style inequality which is of independent interest, and proven in Appendix A.

2 Preliminaries; The Algorithm

Statistical Setting We first present our algorithm in the statistical setting, and then show how it can be adjusted to decision-theoretic settings. Consider a "model" $\mathcal{P} = \{p_\theta \mid \theta \in \Theta\}$ of densities on a space $\mathcal{Z}$ relative to some fixed measure μ. The densities may be, but are not necessarily, probability densities or mass functions: we only require that for all $z \in \mathcal{Z}$, $p_\theta(z) \geq 0$, and $\int_{\mathcal{Z}} p_\theta d\mu < \infty$. We extend p_θ to sequences $z^n = z_1, \ldots, z_n$ of n outcomes by

$p_\theta(z^n) = \prod_{i=1}^n p_\theta(z_i)$. There are no restrictions on the structure of Θ; thus $\mathcal{P}$ may very well be a 'nonparametric' set such as, say, the set of all Gaussian mixtures on $\mathcal{Z}$ with a countable number of components. Often we are interested in estimating a *conditional* probability density. In that case, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and $p_\theta(z)$ abbreviates $p_\theta(y \mid x)$, the conditional density of y given x, and our requirement becomes that for all $x \in \mathcal{X}$, $\int_\mathcal{Y} p_\theta(y \mid x) d\mu_x < \infty$ for some underlying measure μ_x. The abbreviation $z \equiv y \mid x$ is unusual, but in our case harmless, and greatly simplifies notation.

An *estimator* is a function $\breve{\nu} : \bigcup_{n=1}^\infty \mathcal{Z}^n \rightarrow \Theta$ where the function evaluated at z^n is denoted $\breve{\nu} | \, z^n$. If Z^n has a distribution P^*, $\breve{\nu}$ becomes a random variable and we omit the argument '$| \; Z^n$' if it is clear from the context. A *randomized* estimator is a function $\breve{W} : \bigcup_{n=1}^\infty \mathcal{Z}^n \rightarrow \text{dist}(\Theta)$, where $\text{dist}(\Theta)$ is the set of all distributions on Θ. We write $\breve{W} \mid Z^n$ for the estimate for data Z^n. Following Zhang (2006a,b), for any prior Π with density π relative to some underlying measure ρ, we define the *generalized Bayesian posterior*, denoted as $\Pi \mid Z^n, \eta$, as the distribution on Θ with density

$$\pi(\theta \mid z^n, \eta) := \frac{p_\theta^\eta(z^n)\pi(\theta)}{\int_\Theta p_\theta^\eta(z^n)\pi(\theta)\rho(d\theta)} = \frac{p_\theta^\eta(z^n)\pi(\theta)}{\mathbf{E}_{\theta\sim\Pi}[p_\theta^\eta(z^n)]}. \tag{1}$$

We can think of the generalized Bayesian posterior as a randomized estimator. For a randomized estimator $\breve{W}$ and a sample Z^n, we define the corresponding (randomized) Cesàro-averaged estimator as $\text{CES}(\breve{W}; Z^n) := n^{-1} \sum_{i=1}^n \breve{W} | \, Z^i$. We are now ready to present the safe Bayesian algorithm.

Algorithm 1: The Safe Bayesian Algorithm. In the DTOL and statistical learning interpretation, log-loss in the fifth-to-last line is replaced by the loss of interest $\ell_{\Pi|z^{i-1},\eta}(z_i)$. The definition of $\kappa_{\max}$ is explained below (8).

```
Input: data $z_1, \ldots, z_n$, model $\{p_\theta \mid \theta \in \Theta\}$, prior $\Pi$ on $\Theta$.
Output: Distribution on $\Theta$.
$\kappa_{\max} := \lceil \log_2(2\sqrt{n} \ln V) \rceil$ with $V$ as in (3),
$\mathcal{S}_n := \{1, 2^{-1}, 2^{-2}, 2^{-3}, \ldots, 2^{-\kappa_{\max}}\}$ ;
for all $\eta \in \mathcal{S}_n$ do
    $s_\eta := 0$ ;
    for $i = 1 \ldots n$ do
        Compute generalized Bayes posterior $\Pi(\cdot \mid z^{i-1}, \eta)$ with learning rate $\eta$;
        Calculate "posterior expected loss" of predicting actual next outcome:
        $r := E_{\theta\sim\Pi|z^{i-1},\eta}[-\ln p_\theta(z_i)]$ [$= \ell_{\mathbf{\Pi|z^{i-1},\eta}}(\mathbf{z_i})$] ; set  $s_\eta := s_\eta + r$;
    end
end
Choose $\hat{\eta} = \arg\min_{\eta\in\mathcal{S}_n}\{s_\eta\}$ (if min achieved for several $\eta \in \mathcal{S}_n$, pick largest) ;
Output distribution $\breve{W}_{\text{SAFE}} \mid Z^n := \text{CES}(\Pi \mid \hat{\eta}; Z^n) = n^{-1}\sum_{i=1}^n \Pi(\cdot \mid z^i, \hat{\eta})$.
```

The algorithm implements a particular randomized estimator: it picks the $\hat{\eta}$ for which the cumulative log-loss of sequentially predicting by *randomizing* according to the posterior ("Gibbs sampling") is minimized (this is different from standard Bayesian prediction, which *mixes* rather than randomizes). It then outputs the corresponding Cesàro estimator. The use of randomization makes $\hat{\eta}$ very different from a standard 'empirical Bayes' estimate — see Example 4.

DTOL Setting We consider a variation of the original decision-theoretic online (DTOL) setting (Freund and Schapire, 1997) along the lines of (Zhang, 2006a). Let $\mathcal{A}$ be a set of *actions*, where each $a \in \mathcal{A}$ is identified by its *loss* $\ell_a : \mathcal{Z} \to \mathbb{R}$. Thus the loss of action a on outcome z is $\ell_a(z)$. We let $\Theta \subset \mathcal{A}$ be a subset of actions whose losses $\ell_\theta(z_i)$ can be observed at each time point i. As in the original DTOL setting, the learner may not have access to z^{i-1} directly. We assume that the learner is allowed to *randomize*, i.e. for any distribution W in $\mathcal{A}$, all $z \in \mathcal{Z}$ we define

$$\ell_W(z) := \mathbf{E}_{a \sim W}[\ell_a(z)], \tag{2}$$

and we assume that for each such W, the learner is allowed to play an action $a_W \in \mathcal{A}$ with, for all $z \in \mathcal{Z}$, $\ell_{a_W}(z) \leq \ell_W(z)$. This is achieved either automatically (e.g. with convex loss functions defined on convex $\mathcal{A}$, such that for each W an a_W trivially exists) or by definition; e.g. in the PAC-Bayesian literature, it is usually assumed that the learner is allowed to play a randomized action W and is satisfied by evaluating its performance 'on average' (Catoni, 2007).

To apply our algorithm in the DTOL setting, we define pseudo-probabilities $p_a(z) := \exp(-\ell_a(z))$ in the usual manner, for each $a \in \mathcal{A}$, so that $-\ln p_a(z) = \ell_a(z)$, as already indicated in the fifth-to-last-line in Algorithm 1. Readers familiar with the HEDGE-algorithm (Freund and Schapire, 1997, Chaudhuri et al., 2009) will notice that the safe Bayesian algorithm really just runs Hedge at different learning rates η, picking the $\hat{\eta}$ that minimizes cumulative loss with hindsight, and then makes a Cesàro-averaged prediction of the n previous Hedge predictions with this loss. Note however that, while the algorithm employs an on-line learning method, our aim is to prove bounds on its batch behaviour after observing z^n (Theorem 2).

Statistical Learning Setting A special case of the decision-theoretic setting is standard statistical learning in which $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, each θ is a function $\theta : \mathcal{X} \to \mathcal{Y}'$ and $\ell_\theta = \textsc{loss}(Y, \theta(X))$ where $\textsc{loss} : \mathcal{Y} \times \mathcal{Y}' \to \mathbb{R}$ is some loss function, e.g. the 0/1-loss in the classification setting with $\mathcal{Y} = \mathcal{Y}' = \{0, 1\}$, and $\textsc{loss} : \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}$ given by $\textsc{loss}(y, \hat{y}) := |y - \hat{y}|$.

Condition on $\mathcal{P}/\Theta$ Throughout this paper we assume the model $\mathcal{P}$ satisfies the following condition. Let

$$U(\mathcal{P}) := \sup_{z \in \mathcal{Z}} \sup_{p, p' \in \mathcal{P}} \frac{p(z)}{p'(z)} \quad \text{and} \quad V = V(\mathcal{P}) = 2U(\mathcal{P}). \tag{3}$$

If $\mathcal{P}$ is clear from the context we write V rather than $V(\mathcal{P})$ (the reason for distinguishing between V and U is just for notational convenience in stating our results later; it is due to the factor 2 in (4) below). We always assume that the ratio in (3) is well-defined for all $z \in \mathcal{Z}$ and that $1 < V < \infty$. We may think of $U(\mathcal{P})$ as the maximum ratio between the density of z (or $y \mid x$)

assigned by different $p \in \mathcal{P}$. In the DTOL setting with bounded loss functions like 0/1-loss, this condition is automatically satisfied with $V \leq \exp(L_{\max})$ with $L_{\max} = \sup_{z \in \mathcal{Z}, \theta, \theta' \in \Theta} (\ell_\theta(z) - \ell_{\theta'}(z))$. Yet in general density estimation and unbounded loss settings, this is currently a serious restriction on our results.

3 The Critical $\boldsymbol{\eta}_{\text{CRIT}}$ — Extended Introduction

From now on, we assume the random variables Z and $Z_1, Z_2, \ldots, Z^n$ to be i.i.d. $\sim P^*$, i.e. each outcome Z_i has the same distribution as Z. We denote expectation under P^* by $\mathbf{E}^*$. Let $\mathcal{P}$ be the learner's model. Let $D(P\|q)$ denote the KL divergence between a distribution P and a distribution with density q (possibly defective, i.e. $\int_{\mathcal{Z}} q(z) d\mu \neq 1$). In Appendix A we show that that the set of *best-approximating densities*,

$$\mathcal{Q} := \{q : \inf_{p \in \mathcal{P}} D(P^* \| p) = D(P^* \| q) \ , \ U(\mathcal{P} \cup \{q\}) \leq 2U(\mathcal{P})\}, \tag{4}$$

is not empty, although it may not be contained in $\mathcal{P}$ but only in its (appropriately defined) closure (U is as in (3)).

Our Goal We focus on the statistical setting; for the DTOL setting, see Example 2 below. In case $\mathcal{P}$ is a standard probability model (all densities integrate to 1) and $\inf_{p \in \mathcal{P}} D(P^* \| p)$ is nonzero, we say that the model $\mathcal{P}$ is *misspecified* (or simply: “wrong”). Our goal is to show that even in this case, the safe Bayesian algorithm outputs an estimator $\breve{W}_{\text{SAFE}}$ that “converges” quickly to $\mathcal{Q}$, in a sense we now make precise. For any two (conditional) densities p and p', we define the *generalized KL (Kullback-Leibler) divergence* (already introduced in the original Kullback and Leibler (1951)!) relative to P^* as

$$D^*(p' \| p) := \mathbf{E}^*_Z[-\ln p(Z) + \ln p'(Z)] = D(P^* \| p) - D(P^* \| p'). \tag{5}$$

Note that, for a best-approximating q as in (4), $D^*(q \| p) \geq 0$ for all $p \in \mathcal{P}$. Theorem 2 below shows that for some $q \in \mathcal{Q}$, $\mathbf{E}_{\theta \sim \breve{W}_{\text{SAFE}} | Z^n}[D^*(q \| p_\theta)]$ converges to 0 in expectation as $n \to \infty$ at certain rates. Since trivially, for all $q, q' \in \mathcal{Q}$, all $p \in \mathcal{P}$, $D^*(q \| p) = D^*(q' \| p)$, this means that such convergence takes place simultaneously for *all* $q \in \mathcal{Q}$. Hence, from now on, for ease of exposition, we fix a particular such q and present all results in terms of that q. Since $D^*(q \| p_\theta) \geq 0$ for all $\theta \in \Theta$, this convergence implies that, at large n, $\breve{W}_{\text{SAFE}}$ puts nearly all its mass on Θ with small $D^*(q \| p_\theta)$; in this sense, Theorem 2 shows that $\breve{W}_{\text{SAFE}}$ *concentrates.* To make this precise, we must first define the *critical learning rate.*

The Critical Learning Rate In the well-specified case, in which P^* has density p^* and we must have $q = p^* \in \mathcal{Q}$, we trivially have that, for $\eta = 1$, for all $p \in \mathcal{P}$:

$$A_\eta(q \| p) := \ \mathbf{E}^*_Z \left[\left(\frac{p(Z)}{q(Z)} \right)^\eta \right] \leq 1, \tag{6}$$

as is seen by writing out the expectation in full and substituting q by p^*. Classical theorems on two-part MDL inference for the well-specified case (Barron and Cover, 1991, Zhang, 2006a, Grünwald, 2007) invariably make use of (6) at some point in the proofs; so do classical results on Bayesian consistency (Doob, 1949), in which (6) is used to establish that $\{p(Z^n)/q(Z^n)\}_{n=1,2\ldots}$ is a martingale. It can be shown (Li, 1999, Kleijn and van der Vaart, 2006) that (6) still holds for $\eta = 1$ if $\mathcal{P}$ is *convex* (Figure 1 in Section 4 will make clear that convexity plays a role here). This is the fundamental reason why standard MDL and Bayesian convergence bounds still hold in that setting. If (6) does not hold for $\eta = 1$ then MDL and Bayes may not converge — see Example 3 below. Luckily, for many types of $\mathcal{P}$, one can still show that (6) holds for some $\eta < 1$. In that case, the standard MDL and Bayesian convergence proofs still go through if the standard posterior is replaced by the η-generalized posterior, leading to results like (11) below. Thus it makes sense to define the *critical exponent* η_{CRIT} as the largest value of η such that, for all $p \in \mathcal{P}$, (6) holds. It is useful to extend the idea slightly so that, for $u \geq 0$, $\eta_{\text{CRIT}}(u)$ is the "critical exponent with slack u/n"; $\eta_{\text{CRIT}}(0)$ is just the critical value as defined before:

$$\eta_{\text{CRIT}}(u) := \sup \left\{ \eta \leq 1 \;:\; \text{for all } p \in \mathcal{P}, \quad \ln \mathbf{E}^*_Z \left[\left(\frac{p(Z)}{q(Z)} \right)^{\eta} \right] \leq \frac{u}{n} \right\}. \qquad (7)$$

This definition implicitly depends on q and n. Clearly $\eta_{\text{CRIT}}(u)$ is increasing in u. By differentiation to η as in (Grünwald, 2011) it follows that also for all $0 < \eta \leq \eta_{\text{CRIT}}(u)$, all $p \in \mathcal{P}$, $\ln \mathbf{E}^*_Z \left[\left(\frac{p(Z)}{q(Z)} \right)^{\eta} \right] \leq \frac{u}{n}$. In case $\mathcal{Q}$ is not a singleton, we define $\eta_{\text{CRIT}}(u)$ as (7) for the $q \in \mathcal{Q}$ that maximizes it for the given u.

How small can η_{CRIT} become? Let V be as in (3). (Grünwald, 2011, Lemma 1) shows that, for all $P^*, \mathcal{P}$, all $0 \leq u \leq n$,

$$\eta_{\text{CRIT}}(u) \geq \eta_{\text{MIN}}(u), \text{ where } \eta_{\text{MIN}}(u) := \tfrac{1}{2 \ln V} \sqrt{\tfrac{u}{n}}. \qquad (8)$$

To get good bounds on the behaviour of the Safe Bayesian algorithm as in Theorem 2 we need to be able to use an η close to $\eta_{\text{CRIT}}(u)$ for a value of $u \geq 0$ that optimizes the bound in Theorem 2. It can be seen that restricting u to be ≥ 1 does not seriously affect the bound, which explains why, in the definition of $\mathcal{S}_n$ in the safe Bayesian algorithm, we could safely restrict ourselves to η no smaller than $O(1/(2 \ln V \sqrt{n}))$. In favourable cases though, $\eta_{\text{CRIT}}(u)$ will be larger than $\eta_{\text{MIN}}(u)$. We shall now see that this leads to faster convergence rates.

Existing Results that we will Extend We define the generalized Bayesian marginal distribution as $p_{\text{Bayes}}(z^n \mid \eta) := \mathbf{E}_{\theta \sim \Pi} [p^{\eta}_{\theta}(Z^n)]$ and the predictive distribution as $p_{\text{Bayes}}(z_i \mid z^{i-1}, \eta) := p_{\text{Bayes}}(z^i \mid \eta) / p_{\text{Bayes}}(z^{i-1} \mid \eta)$. For $\eta = 1$, these are the standard Bayesian marginal/predictive distributions. By the familiar Bayesian telescoping using (1), p_{Bayes} can be written as product of the generalized posterior predictive distributions:

$$p_{\text{Bayes}}(z^n \mid \eta) = \prod_{i=1}^n \tfrac{p_{\text{Bayes}}(z^i \mid \eta)}{p_{\text{Bayes}}(z^{i-1} \mid \eta)} = \prod_{i=1}^n \mathbf{E}_{\theta \sim \Pi \mid Z^{i-1}, \eta} \left[p^{\eta}_{\theta}(z_i) \right]. \qquad (9)$$

We also define the *Bayesian redundancy* as

$$\textsc{bayes-red}_n(\eta) := \tfrac{1}{\eta}\mathbf{E}^*_{Z^n}\left[-\ln\tfrac{p_{\text{Bayes}}(Z^n|\eta)}{q^\eta(Z^n)}\right] = \tfrac{1}{\eta}\mathbf{E}^*_{Z^n}\left[\textstyle\sum_{i=1}^n -\ln\tfrac{p_{\text{Bayes}}(Z_i|Z^{i-1}\eta)}{q^\eta(Z_i)}\right] \tag{10}$$

For $\eta = 1$, the Bayes redundancy is the expected codelength difference between coding (log-loss prediction) by the code induced by $p_{\text{Bayes}} \mid \eta$ and coding by the code induced by q. This quantity is indeed called the (relative) "redundancy" of the Bayesian mixture code in information theory, see e.g. (Takeuchi and Barron, 1998). We can also give a precise codelength interpretation for $\eta < 1$ via the 'entropification' construction as in Grünwald (2011), but because of space constraints will not do so here. We now informally summarize a central result in MDL and PAC-Bayesian inference in terms of BAYES-RED: for all $0 < \eta < \eta_{\text{CRIT}}(u)$, for some constant C_η depending on η, we have

$$\mathbf{E}^*_{Z^n}\mathbf{E}_{\theta\sim\Pi|Z^n,\eta}\left[D^*(q\|p_\theta)\right] \leq \tfrac{C_\eta}{n}\cdot\textsc{bayes-red}_n(\eta) + R_u, \tag{11}$$

where R_u is a remainder term that becomes negligible for small enough $u \geq 0$. In the remainder of this informal section, we assume that we have chosen u small enough and ignore this term, as well as other precise conditions needed for (11) to hold (R_u will return in the formal statement of our results). (11) is the generic formulation of the result. Variations of (11) are presented by, among others, Zhang (2006a,b), Barron and Cover (1991), Li (1999), Audibert (2004), Catoni (2007). The importance of (11) is that *in practical settings* BAYES-RED$_n(\eta)$ *grows sublinearly and then (11) implies that (a) the generalized posterior concentrates and (b) leads to asymptotically optimal approximations to q in KL divergence.*

Example 1. **[MDL formulation]** A simple rewriting of the redundancy as in Zhang (2006b) shows that

$$\textsc{bayes-red}_n(\eta) = \mathbf{E}^*_{Z^n}\left[\mathop{\mathbf{E}}_{\theta\sim\Pi|Z^n,\eta}\left[-\ln\frac{p_\theta(Z^n)}{q(Z^n)}\right] + \frac{1}{\eta}D(\,(\Pi\mid Z^n,\eta)\,\|\,(\Pi\mid\eta)\,)\right] \tag{12}$$

where $D(W\|V) = \int w(\theta)\log(w(\theta)/v(\theta))\rho(d\theta)$ denotes standard KL divergence between distributions with densities W and V respectively. To verify (12), simply replace D by its definition and simplify. If Π has countable support $\Theta' \subset \Theta$ then, irrespective of model correctness, using that for all θ_0, all z^n, $p_{\text{Bayes}}(z^n \mid \eta) = \sum\pi(\theta)p^\eta_\theta(z^n) \geq \pi(\theta_0)p^\eta_{\theta_0}(z^n)$, we have the familiar

$$\textsc{bayes-red}_n(\eta) \leq \mathbf{E}^*_{Z^n}\left[\min_{\theta\in\Theta'}\ \left\{\ -\ln\tfrac{p_\theta(Z^n)}{q(Z^n)} + \tfrac{-\ln\pi(\theta)}{\eta}\ \right\}\ \right]. \tag{13}$$

If $q = p_{\tilde\theta}$ for some $\tilde\theta \in \Theta'$, this becomes BAYES-RED$_n(\eta) \leq -\ln\pi(\tilde\theta)/\eta$, showing that then prediction by p_{Bayes} stays within $O(1)$ of the best-approximating q.

Example 2. **[Statistical Learning]** Now $z = (x, y)$, $\ell_\theta(z) = \textsc{loss}(y, \theta(x))$, we define RISK$(\theta)$ to be the expected loss of θ, i.e. RISK$(\theta) := \mathbf{E}^*_Z[\ell_\theta(Z)]$, extended to RISK$(W)$ as in (2). Let RISK$_{\text{emp}}(W) = n^{-1}\sum_{i=1}^n \ell_W(Z_i)$ be the empirical risk of distribution W. Let $\tilde{\theta}$ be any optimal action within Θ, i.e. $\text{RISK}(\tilde{\theta}) = \min_{\theta \in \Theta} \text{RISK}(\theta)$. Then using $\ell_\theta = -\ln p_\theta$, (12) can be further rewritten as

$$\tfrac{1}{n}\text{BAYES-RED}_n(\eta) = \mathbf{E}^*_{Z^n}\left[\text{RISK}_{\text{emp}}(\Pi \mid Z^n, \eta)\right] - \text{RISK}(\tilde{\theta}) + \eta^{-1}\mathbf{E}^*_{Z^n}[D(\cdot \| \cdot)]$$

and (11) now expresses that , with $R := \mathbf{E}^*_{Z^n}\left[\text{RISK}_{\text{emp}}(\Pi \mid Z^n, \eta)\right] - \text{RISK}(\tilde{\theta})$,

$$\mathbf{E}^*_{Z^n}[\text{RISK}(\Pi \mid Z^n, \eta)] - \text{RISK}(\tilde{\theta}) \leq C \cdot R + \tfrac{C}{n\eta}\mathbf{E}^*_{Z^n}[D(\ (\Pi \mid Z^n, \eta) \parallel (\Pi \mid \eta)\ )],$$

a familiar equation from the PAC-Bayesian literature: the relative risk is bounded by the empirical risk difference plus a KL-divergence penalty term. Analogous results hold in probability rather than in expectation (in many of the PAC-Bayesian literature, only in-probability results are given; Zhang (2006a, 2006b) gives both in-probability and in-expectation results).

The bounds that can be obtained via (11) are often minimax optimal. For example, if the model is correct then $\eta_{\text{CRIT}}(0) = 1$, so we can take $u = 0$. For that case Barron and Cover (1991) already showed that with appropriate choices of prior $\text{BAYES-RED}_n(1)$, (or rather its upper bound (13)) is so small that (11) leads to the optimal convergence rates in a number of nonparametric settings; Zhang (2006a) extends this to parametric models $\mathcal{P}$. If we consider 0/1-loss and a countable set of classifiers Θ, then, as is well-known, the worst-case risk obtainable by any procedure is $O((-\ln \pi(\tilde{\theta})/n)^{1/2})$ and as shown by Grünwald (2011), this is indeed the bound we get from (11) if u is chosen appropriately. Many other examples can be found in (Zhang, 2006a,b).

The key point for us is that (11) only holds for $\eta < \eta_{\text{CRIT}}(u)$; but $\eta_{\text{CRIT}}(u)$ depends on the true distribution and it is not clear how to find it. Our Theorem 1 combined with Theorem 2 imply via Corollary 1 that the safe Bayesian algorithm $\breve{W}_{\text{SAFE}}$ performs at least as well as the Bayesian posterior randomized estimator $\Pi \mid \eta$ with $\eta = \eta_{\text{CRIT}}(u)/4$. Since $\text{BAYES-RED}_n(\eta)/n$ has a bounded nonnegative derivative (as is straightforward to show), this leads to bounds that are within a constant factor of the best bound that can be obtained for any $\eta \leq \eta_{\text{CRIT}}(u)$.

In fact, Theorem 2 only shows that $\breve{W}_{\text{SAFE}}$ satisfies (11) plus an additional penalty MIX-GAP_n, which measures how much is lost in terms of cumulative log-loss by randomizing rather than mixing. Theorem 1 below shows that, for $\eta \leq \eta_{\text{CRIT}}(u)/2$, this extra penalty is sufficiently small to get the desired bound.

4 First Result: Randomizing vs. Mixing

Define the *Gibbs redundancy* as

$$\text{GIBBS-RED}_n(\eta) = \mathbf{E}^*_{Z^n}\left[\textstyle\sum_{i=1}^n \mathbf{E}_{\theta \sim \Pi \mid Z^{i-1}, \eta}\left[-\ln \frac{p_\theta(Z_i)}{q(Z_i)}\right]\right].$$

and note that, by Jensen's inequality and (10), we always have $\text{BAYES-RED}_n(\eta) \leq \text{GIBBS-RED}_n(\eta)$. The following theorem shows that, if η is sufficiently subcritical, then the reverse essentially holds as well:

Theorem 1. *Let $\eta_{\text{CRIT}}(u)$ be defined as in (7). For $0 < \eta \leq \eta_{\text{CRIT}}/2$, we have:*

$$\text{GIBBS-RED}_n(\eta) \leq C_{2\eta}\text{BAYES-RED}_n(\eta) + (C_{2\eta} - 1)\tfrac{u}{\eta}, \tag{14}$$

for a constant $C_\eta \leq 2 + 2\eta \ln V$ (so $C_{2\eta} \leq 2 + 4\eta V$) with V as in (3).

The theorem thus expresses that, in terms of log-loss, if $\eta \leq \eta_{\text{CRIT}}(u)/2$ then sequential prediction by posterior randomization is not much worse in expectation than sequential prediction by the standard Bayes predictive distribution, i.e. by mixing rather than randomizing. To explore this further, we define the *mixability gap* of a randomized estimator $\breve{W}$ as

$$\text{MIX-GAP}_n(\eta, \breve{W}) := \mathbf{E}^*_{Z^n}\left[\sum_{i=1}^n \mathbf{E}_{\theta\sim\breve{W}|Z^{i-1}}[-\ln p_\theta(Z_i)] + \tfrac{1}{\eta}\ln p_{\text{Bayes}}(Z^n \mid \eta)\right]$$

The mixability gap for the Bayesian posterior can be rewritten as:

$$\text{MIX-GAP}(\eta,\ (\Pi|\eta)\ ) = \text{GIBBS-RED}_n(\eta) - \text{BAYES-RED}_n(\eta) \geq 0. \tag{15}$$

In the information-theoretic interpretation, MIX-GAP_n is the expected amount of additional bits (additional log-loss), normalized relative to η, incurred by predicting by randomizing according to the posterior rather than by p_{Bayes}, which first mixes using the posterior and then predicts using the resulting distribution. With these definitions, (14) can be rewritten as $(\text{MIX-GAP}_n(\eta, \Pi \mid \eta) + \text{BAYES-RED}_n(\eta))$ $\leq C_{2\eta}(\text{BAYES-RED}_n(\eta) + (C_{2\eta} - 1)u/\eta$, i.e.

$$\text{MIX-GAP}_n(\eta, \Pi \mid \eta) \leq (C_{2\eta} - 1)\left(\text{BAYES-RED}_n(\eta) + \tfrac{u}{\eta}\right). \tag{16}$$

Hence, for $\eta \leq \eta_{\text{CRIT}}(u)/2$, the excess loss of randomizing rather than mixing is of the same order as the excess loss of mixing rather than predicting with q.

Example 3. [**Bayesian misspecification**] For simplicity consider Π with countable support. As shown by Grünwald and Langford (2007), if the model is incorrect, in some cases with $\eta_{\text{CRIT}}(0) \ll 1$, the standard Bayesian posterior (based on $\eta = 1$) puts, P^*-almost surely, nearly all of its mass on a set of 'bad' distributions p', all with arbitrarily large $D^*(q\|p')$ at all large n. Yet (13) shows that the redundancy, and hence the cumulative log-loss risk of standard Bayesian prediction (with $\eta = 1$) must still be small (see Example 5); this is possible because Bayes then *mixes* various 'bad' but very different $p' \in \mathcal{P}$ into a single 'good' predictive distribution $p_{\text{Bayes}}(Z_i \mid Z^{i-1}, \eta) \notin \mathcal{P}$; see Figure 1. If this happens[1] for many i between 1 and n, then by definition $\text{MIX-GAP}_n(\eta, \Pi \mid \eta)$ becomes extremely large. Theorem 1 shows that, if we set $\eta \leq \eta_{\text{CRIT}}(u)/2$, then this will not happen: the posterior $\Pi \mid Z^n, \eta$ will concentrate in the sense that if we sample from it, we will tend to draw a distribution p with $D^*(q\|p)$ close to 0, for all $q \in \mathcal{Q}$. Even if $\mathcal{Q}$ is nonsingleton this is fundamentally different from choosing $\eta = 1 \gg \eta_{\text{CRIT}}(0)/2$, in which case the posterior puts almost all of its mass on distributions p' with $D^*(q\|p')$ large for all $q \in \mathcal{Q}$. Example 5 explains why posterior concentration is so important.

[1] In the GL examples, the set of distributions over which the posterior mixes changes with sample size i, but they always remain 'bad', yet the resulting predictive distribution always remains 'good', i.e. $D^*(q\|p_{\text{Bayes}}(Z_i \mid Z^{i-1}, \eta))$ remains small.

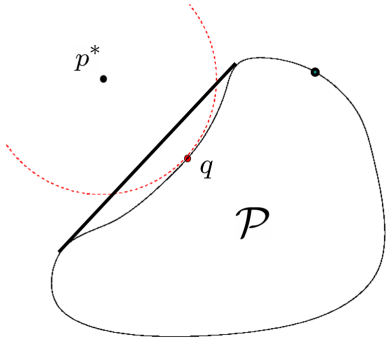

Fig. 1. Mixing vs. Randomizing: a mixture (e.g. the Bayes predictive distribution) that puts substantial mass on the two endpoints of the line segment, is closer to p^* than q, the best approximation of p^* within $\mathcal{P}$. This can only happen if q is not in the convex hull of $\mathcal{P}$. The picture is an idealization: in the GL examples the posterior mixes not just two but many 'bad' (i.e., far from p^*) distributions, $\mathcal{P}$ is not 2-dimensional parametric and the geometry is not Euclidean but determined by the KL divergence.

5 Main Result and Its Applications

We have seen that the learner would like to infer a $\hat{\eta}$ from the data that works at least as well as the unknown $\eta_{\text{CRIT}}(u)/2$. The following theorem shows that the safe Bayesian algorithm achieves this. Let $\mathcal{S}_n$ be defined as in Algorithm 1, and let $\{\breve{W}_\eta\}$ for $\eta \in \mathcal{S}_n$ represent a set of randomized estimators for Θ, one for each η. We let $\hat{\eta}$ be the "maximum likelihood" estimate of η, i.e. $\hat{\eta} \mid Z^n = \arg\min_{\eta \in \mathcal{S}_n} \sum_{i=1}^n \mathbf{E}_{\theta \sim W \mid Z^{i-1}, \eta} [-\ln p_\theta(Z_i)]$.

Theorem 2. *Let $C_{2\eta}$ be as in Theorem 1. For $\eta \in \mathcal{S}_n$ with $\eta \leq \eta_{\text{CRIT}}(u)/2$, we have:*

$$\begin{aligned}
&\mathbf{E}^*_{Z^n} \mathbf{E}_{\theta \sim \text{CES}(\breve{W}_{\hat{\eta} \mid Z^n}; Z^n)} \left[D^*(q \| p_\theta) \right] \leq \\
&\frac{C_{2\eta}}{n} \mathbf{E}^*_{Z^n} \left[\sum_{i=1}^n \mathbf{E}_{\theta \sim \breve{W}_\eta \mid Z^{i-1}} \left[-\ln \frac{p_\theta(Z_i)}{q(Z_i)} \right] + \frac{u + O(\ln \ln n)}{\eta} \right] = \\
&\frac{C_{2\eta}}{n} \left(\text{MIX-GAP}_n(\eta, \breve{W}_\eta) + \text{BAYES-RED}_n(\eta) + \frac{u + O(\ln \ln n)}{\eta} \right).
\end{aligned} \tag{17}$$

The theorem works for any $\breve{W}_\eta$, but to get good bounds the mixability gap of $\breve{W}_\eta$ must be small. Theorem 1 tells us that it will be small if we use $\breve{W}_\eta \mid Z^n := (\Pi \mid Z^n, \eta)$, i.e. we randomize according to the posterior. If we plug this choice into (17) and rewrite the right-hand side using Theorem 1 (see (16)) and the fact that $\text{BAYES-RED}_n(\eta)$ is decreasing in η and $\eta_{\text{CRIT}}(u) \geq \eta_{\text{MIN}}(u)$ as in (8), then:

Corollary 1. *The Safe Bayesian algorithm satisfies, for $\eta \leq \eta_{\text{CRIT}}(u)/4$:*

$$\mathbf{E}^*_{Z^n} \mathbf{E}^*_{\theta \sim \breve{W}_{\text{SAFE}} \mid Z^n} \left[D^*(q \| p_\theta) \right] \leq \frac{C^2_{2\eta}}{n} \left(\text{BAYES-RED}_n(\eta) + \frac{u + O(\ln \ln n)}{\eta} \right). \tag{18}$$

Note that $C_{2\eta}$ has become $C^2_{2\eta}$. We got rid of the requirement that $\eta \in \mathcal{S}_n$ by using that $\text{BAYES-RED}_n(\eta)$ is decreasing in η, so that (18) must hold for every η smaller than the largest $\eta_{\max} \in \mathcal{S}_n$ with $\eta_{\max} \leq \eta_{\text{CRIT}}(u)/2$. Note that $\eta_{\max}$ may be arbitrarily close to $\eta_{\text{CRIT}}(u)/4$ rather than $\eta_{\text{CRIT}}(u)/2$.

While the bound is thus in terms of $\eta \leq \eta_{\text{CRIT}}(u)/4$, it is conceivable that the algorithm chooses a larger $\hat{\eta}$, possibly even with $\hat{\eta} \gg \eta_{\text{CRIT}}(u)$. Thus, to be fully precise, we cannot claim that we "learn" the optimal learning rate, but only that we learn to behave as well as if we would know the optimal learning rate. The second line of (17) indicates that the randomized η-posterior in Algorithm 1 may in principle be replaced by other estimators that approximate $p_{\text{Bayes}} \mid \eta$ reasonably well, such as e.g. the Bayesian MAP estimator with prior $w(\theta)^{1/\eta}$.

Example 4. **[Safe vs. Empirical Bayes]** The Safe Bayesian algorithm chooses $\hat{\eta}$ that minimizes a cumulative log-loss and hence maximizes a likelihood[2]. Indeed, if we interchanged expectation and logarithm in the definition of r in Algorithm 1, then we would mix rather than randomize and by (9), $\hat{\eta}$ would become the *empirical Bayes estimate* of η. Now for $\eta > \eta_{\text{CRIT}}(u)$, we may be in the situation of Figure 1 where our Bayesian predictive distribution achieves small cumulative log loss by mixing, at many sample sizes, bad distributions into a good one. Empirical Bayes will tend to pick such an unwanted η, and indeed, it was already shown in GL that it does not solve the Bayesian inconsistencies noted there. Similarly, a Bayesian may want to treat η as a nuisance parameter and integrate it out, but this approach fails for the same reason. Intuitively, by randomizing rather than mixing, the safe Bayesian estimator is guarded from picking such an overly large η, whereas by Theorem 1, it is guarded from picking an η much smaller than $\eta_{\text{CRIT}}(u)$.

Example 5. **[Bayesian Misspecification, Cont.]** The examples considered by GL are based on P^* and countable $\mathcal{P}$ such that $0 < \eta_{\text{CRIT}}(0) \ll 1$ and a prior $\pi(q) > 0$ on the best-approximating q. Using (13), Corollary 1 thus bounds the convergence rate of $\breve{W}_{\text{SAFE}}$ as $O((\ln \ln n)/n)$, only a factor $\ln \ln n$ worse compared to (11), which is the best known bound when $\eta_{\text{CRIT}}(0)$ is known in advance. Thus, the inconsistency for $\eta = 1$ goes away. Now one may wonder why one should not just, given sample Z^n, directly use the standard Bayes predictive distribution $p_{\text{Bayes}}(Z_{n+1} \mid Z^n, \eta)$ with $\eta = 1$ to make predictions about Z_{n+1}? By (10) and (13), the cumulative expected log-loss risk of this procedure should be bounded by $-\ln \pi(q)$, indicating a risk smaller than $O(1/n)$ at most n. If one is only interested in log-loss, this can indeed be done, and there is indeed no good reason to take $\eta < 1$. But in many other cases, there is a very good reason to take a smaller $\hat{\eta}$ so that Corollary 1 holds. We list three such reasons below, numbered as (I)-(III). Note first that the corollary implies that, for large enough n, the posterior is concentrated (see above Example 1), and the phenomenon of Figure 1 cannot occur (Ex. 3). Now first, (I), for 'nonmixable' loss functions (Vovk, 1990) one cannot mix the predictors θ in a Bayesian way. For example, the examples of GL also have an interpretation in terms of 0/1-loss: they show that, in the statistical learning setting with $p_\theta(y \mid x) \propto \exp(-\text{LOSS}(y, \theta(x)))$, predicting according to the MAP, Gibbs or Bayes classifier based on the posterior $\Pi \mid Z^n, \eta$ for $\eta = 1$

[2] In fact, $\hat{\eta}$ maximizes a "prequential" likelihood, and the algorithm (not its analysis) is a prime instance of Dawid's (1984) prequential approach to statistics, which inspired this work throughout.

lead to predictions that *never* come close to $\tilde{L} = \inf_{\theta \in \Theta} \text{RISK}(\theta)$. Yet Corollary 1 implies (see Example 2) that prediction based on $\breve{W}_{\text{SAFE}}$ does converge to $\tilde{L}$ at rate $O((\ln \ln n)/n)$. But now, in contrast to the log-loss case, predicting with $p_{\text{Bayes}} \mid \eta$ for $\eta = 1$ is not an option, since — as explained at length by GL07 — its predictions, which are mixtures over p_θ as above, are mixtures of exponentiated rather than randomized classifiers and hence do not correspond to feasible actions; rather, they are *pseudo-predictions* in the sense of Vovk (2001). In that case, prediction by $\breve{W}_{\text{SAFE}}$, i.e. using the learning rate $\hat{\eta}$, is presumably the best one can do. Second, (II), even in a standard Bayesian setting where a standard probability model is used rather than the pseudo-probabilities above, one is often interested in a loss function different from log-loss. If the model is chosen with the desired loss function in mind and the posterior is concentrated, one can expect good behaviour with respect to that loss function as well; if the posterior is not concentrated, one can only expect good behaviour in terms of log-loss. We briefly illustrate this for 0/1-loss. *Classification models* $\{p_\theta \mid \theta \in \Theta\}$ are models of conditional distributions, such as, e.g., logistic regression models, that are designed in such a way that, for all θ, if p_θ has small expected log loss, then the classifier induced by p_θ also is guaranteed to have small expected logistic loss, which, under some further conditions, implies a small 0/1-loss (Bartlett et al., 2006) — the logistic loss being a 'proxy' for 0/1-loss. However, as shown in detail by (Grünwald and Langford, 2007, Example 1), even if a mixture of such p_θ has much smaller expected log-loss than its constituents, as in Figure 1 , the expected 0/1-loss of the Bayes classifier corresponding to the mixture may actually be much *worse* in terms of 0/1-loss. Thus, using standard Bayesian inference, we may up with predictions that are good in terms of log-loss but bad in terms of the loss of interest.

Finally, (III), the primary goal in many statistical analyses is not sequential prediction but rather coming up with an insightful, *comprehensible* model that is not too far detached from the truth. In that case one would prefer a posterior pointing to a single, not-too-bad model over a very good mixture of very bad models. For example, Yang (2007) and others use Bayesian inference to learn phylogenetic trees in biology using a stochastic models for DNA mutations. what use would a mixture of bad trees be to a biologist, even if it predicts DNA sequences well in terms of log-loss?

Example 6. **[Probabilistic Setting with Correct or Convex Model $\mathcal{P}$]** Corollary 1 implies that safe Bayesian estimation behaves essentially as well as standard Bayesian estimation if the model is correct, i.e. $\inf_{p \in \mathcal{P}} D(P^* \| p) = D(P^* \| q) = 0$ and $q = p^*$. Then $\eta_{\text{CRIT}}(0) = 1$ and we can take $u = 0$ and $C_{2\eta}^2/\eta = C_2^2/1 \leq (2 + 4 \ln V)^2$ in (18). Zhang (2006a) obtains the same risk bound as (18) for $\Pi \mid \eta$ for any $\eta < 1$, the only difference being that the factor $C_2^2/1$ on the right is replaced by something smaller, and that there is no $O(\ln \ln n/\eta n)$ term. The extra factor incurred by $\breve{W}_{\text{SAFE}}$ may be the inevitable price to pay for not knowing in advance that our model was, in fact, correct, using a procedure that still leads to good results if it is incorrect.

6 Related Work

Grünwald (2011) already proposed an adjustment to 2-part MDL and Bayesian MAP approaches to deal with misspecified (wrong) models. The learning rate was determined in a completely different manner, roughly by measuring how much one can additionally compress the data using a code based on the convex hull of $\mathcal{P}$ rather than $\mathcal{P}$. The resulting procedure is computationally much more demanding than the Safe Bayesian algorithm. Also, it can only be applied to countable $\mathcal{P}$ — a severe restriction — whereas the Safe Bayesian algorithm can be applied to arbitrary $\mathcal{P}$. Finally, the bounds in Grünwald (2011) —although qualitatively similar to the ones presented here — have much larger constant factors ($O(V)$ instead of $O(\ln V)$ with V as in (3) above). On the other hand, the safe Bayesian algorithm has two awkward properties that are absent from the procedure of Grünwald (2011): it is dependent on the order of the data, and it requires Cesàro-averaging. Whether the Cesàro-average is really needed is one of the main questions for future research.

Other related work includes van Erven et al. (2011), who use the mixability gap in a worst-case setting to get a version of the Hedge algorithm that in some cases picks a higher learning rate than standard Hedge and thereby achieves a smaller regret.

Finally, in this short paper we did not mention at all how our work relates to the pioneering Audibert (2004), who proposed a method for learning the learning rate in the special case of PAC-Bayesian classification. Audibert showed that, with his approach, one can achieve fast rates under a variation of the Tsybakov (1999, 2004) margin condition. The same holds for the Safe Bayesian algorithm, as we will show in the full paper.

Acknowledgements Supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This paper benefited from discussions with A. Barron, S. de Rooij, T. van Erven, W. Koolen, A.P. Dawid and some insightful remarks by an anonymous referee.

7 Proofs

Proof of Theorem 1 Apply Lemma 1 below, with $\mathcal{G} = \Theta$, $\nu(\theta) := \theta$ and for all $z^i \in \mathcal{Z}^i$, $f_{\nu(\theta)}(z_i \mid z^{i-1}) := p_\theta(z_i)/q(z_i)$, noting that the Lemma applies for $\eta \leq \eta_{\text{CRIT}}(u)/2$. Rewriting the left-hand side using the definition of GIBBS-RED, the statement is seen to imply Theorem 1.
To prepare for Lemma 1, let $Z, Z_1, Z_2, \ldots Z_n$ be i.i.d. random variables relative to a probability triple (Ω, Σ, P^*). Let $\mathcal{G}$ be a set and, for each $\nu \in \mathcal{G}$, let $f_\nu(\cdot \mid \cdot) : \mathcal{Z} \times \bigcup_{i=0}^{n-1} \mathcal{Z}^i \rightarrow \mathbb{R}^+$ be a measurable function such that for $i \leq n$, $P^*(f_\nu(Z_i \mid Z^{i-1}) > 0) = 1$. The notation $f_\nu(Z_i \mid Z^{i-1})$ is suggestive of our applications, in which f_ν represents a ratio of conditional densities. Define $f_\nu(Z^n) := \prod_{i=1}^n f(Z_i \mid Z^{i-1})$. Let Π be a prior distribution on $\mathcal{G}$ and let $\Pi \mid Z^i, \eta$ be the generalized posterior defined as in (1), with p_θ^η replaced by f_ν^η.

Lemma 1. *Let $C_\eta = 2 + 2\eta \ln V$ and suppose for all $z^n \in \mathcal{Z}^n$, $f_\nu(z^i \mid z^{i-1}) \in [1/V, V]$. For all $\eta > 0$ such that for all $\nu \in \mathcal{G}$, $\ln \mathbf{E}^*_Z[f_\nu^{2\eta}(Z)] \leq u/n$, we have:*

$$\begin{aligned} &\mathbf{E}^*_{Z^n}\left[\textstyle\sum_{i=0}^{n-1} \mathbf{E}_{\nu \sim \Pi \mid Z^i, \eta}[-\ln f_\nu(Z_{i+1} \mid Z^i)]\right] \leq \\ &\tfrac{C_{2\eta}}{\eta} \mathbf{E}^*_{Z^n}\left[-\ln \mathbf{E}_{\nu \sim \Pi} f_\nu^\eta(Z^n)\right] + \tfrac{C_{2\eta}-1}{\eta} u. \end{aligned} \tag{19}$$

Proof.

$$\begin{aligned} &\mathbf{E}^*_{Z^n}\left[\textstyle\sum_{i=0}^{n-1} \mathbf{E}_{\nu \sim \Pi \mid Z^i}[-\ln f_\nu(Z_{i+1} \mid Z^i)]\right] = \\ &\eta^{-1} \textstyle\sum_{i=0}^{n-1} \mathbf{E}^*_{Z^i} \mathbf{E}_{\nu \sim \Pi \mid Z^i}\left[\mathbf{E}^*_{\bar{Z}_{i+1}}[-\ln f_\nu^\eta(\bar{Z}_{i+1} \mid Z^i)]\right] \leq_{\text{(a)}} \\ &\eta^{-1} \textstyle\sum_{i=0}^{n-1} \mathbf{E}^*_{Z^i}\left[C_{2\eta}\left(-\ln \mathbf{E}^*_{\bar{Z}_{i+1}} \mathbf{E}_{\nu \sim \Pi \mid Z^i}[f_\nu^\eta(\bar{Z}_{i+1} \mid Z^i)]\right) + (C_{2\eta} - 1)\left(\tfrac{u}{n}\right)\right] \leq_{\text{(b)}} \\ &\eta^{-1} \textstyle\sum_{i=0}^{n-1} \mathbf{E}^*_{Z^i} \mathbf{E}^*_{Z_{i+1}}\left[C_{2\eta}\left(-\ln \mathbf{E}_{\nu \sim \Pi \mid Z^i}[f_\nu^\eta(Z_{i+1} \mid Z^i)]\right) + (C_{2\eta} - 1)\left(\tfrac{u}{n}\right)\right] = \\ &\tfrac{C_{2\eta}}{\eta} \mathbf{E}^*_{Z^n}\left[\textstyle\sum_{i=0}^{n-1} -\ln \mathbf{E}_{\nu \sim \Pi \mid Z^i}[f_\nu^\eta(Z_{i+1} \mid Z^i)]\right] + \tfrac{C_{2\eta}-1}{\eta} u =_{\text{(c)}} \\ &\tfrac{C_{2\eta}}{\eta} \mathbf{E}^*_{Z^n}\left[-\ln \mathbf{E}_{\nu \sim \Pi} f_\nu^\eta(Z^n)\right] + \tfrac{C_{2\eta}-1}{\eta} u. \end{aligned}$$

(a) follows from Lemma 2 below, applied with T set to the random vector $T = (\nu, \bar{Z}_{i+1})$ and $g((\nu, \bar{Z}_{i+1})) \equiv f_\nu^\eta(\bar{Z}_{i+1} \mid Z^i)$. (b) is Jensen's inequality, and (c) is the telescoping of the Bayesian predictive distribution as in e.g. (9). All other equalities are immediate.

Lemma 2. *Let T be a random vector taking values in some set $\mathcal{T}$. For all measurable functions $g : \mathcal{T} \to [1/V, V]$, all $\eta' > 0, \epsilon \geq 0$ with $\ln \mathbf{E}[g(T)^{2\eta'}] \leq \epsilon$, all $0 < \eta \leq \eta'$: $\mathbf{E}[-\ln g(T)] \leq \frac{C_{2\eta}}{\eta}(-\ln \mathbf{E}[g(T)^\eta]) + \frac{C_{2\eta}-1}{\eta}\epsilon$.*

This lemma slightly extends a result by Barron and Li (1999). It is proved as Proposition 5 in Grünwald (2011) (in different context, but the modification to our setting is immediate).

Proof of Theorem 2 We apply Theorem 3 below in the form (23), with $\mathcal{G}$ set to $\mathcal{S}_n$ in Theorem 2, the deterministic estimator $\breve{\nu}$ set to $\hat{\eta}$, and with $f_{\hat{\nu}_{|Z^n}}(z_i \mid z^{i-1})$ set to $\exp(\mathbf{E}_{\theta \sim \breve{W}_{\hat{\eta}|Z^n} | z^{i-1}}[\eta \ln(p_\theta(z_i)/q(z_i))]$. Here η is just a fixed exponent and $\hat{\eta}$ is the meta-estimator in Theorem 2 indexing the learning rate at which the randomized estimator $\breve{W}_\eta$ of Theorem 2 is applied. Plugging these substitutions into (23) using that $Z_1, Z_2, \ldots$ are i.i.d., we get

$$\begin{aligned} &\mathbf{E}^*_{Z^n}\left[\sum_{i=1}^n -\tfrac{1}{\eta} \ln \mathbf{E}^*_{\bar{Z}_i}\left[f_{\hat{\nu}_{|Z^n}}(\bar{Z}_i \mid Z^{i-1})\right]\right] \leq \\ &\mathbf{E}^*_{Z^n}\left[\sum_{i=1}^n \mathbf{E}_{\theta \sim W_{\hat{\eta}|Z^n} | Z^{i-1}}\left[-\ln \tfrac{p_\theta(Z_i)}{q(Z_i)}\right] + \tfrac{-\ln \pi(\hat{\eta})}{\eta}\right]. \end{aligned}$$

where we also divided both sides by η. We now move the inner expectation on the left-hand side outside of the logarithm by applying Lemma 2 with $T = \bar{Z}_i$, $g^\eta(\bar{Z}_i) = f_{\hat{\nu}_{|Z^n}}(\bar{Z}_i \mid Z^{i-1})$, using our assumption $0 < \eta < \eta_{\text{CRIT}}(u)/2$, which gives

$$\begin{aligned} &\mathbf{E}^*_{Z^n}\left[\textstyle\sum_{i=1}^n \mathbf{E}^*_{\bar{Z}_i} \mathbf{E}^*_{\theta \sim W_{\hat{\eta}|Z^n} | Z^{i-1}}\left[-\ln \tfrac{p_\theta(\bar{Z}_i)}{q(\bar{Z}_i)}\right]\right] \leq \\ &\tfrac{C_{2\eta}}{\eta} \mathbf{E}^*_{Z^n}\left[\textstyle\sum_{i=1}^n -\ln \mathbf{E}^*_{\bar{Z}_i}\left[f_{\hat{\nu}_{|Z^n}}(\bar{Z}_i \mid Z^{i-1})\right] + \tfrac{u}{n}\right]. \end{aligned} \tag{20}$$

Combining the previous two equations, dividing by n and recognizing the inner expectation in the left hand side of (20) to be equal to $\mathbf{E}^*_{\theta \sim \breve{W}_{\hat{\eta} \mid Z^n} \mid Z^{i-1}} D^*(q \| p_\theta)$ gives

$$\begin{aligned} &\tfrac{1}{n} \mathbf{E}^*_{Z^n} \left[\textstyle\sum_{i=1}^n \mathbf{E}^*_{\theta \sim \breve{W}_{\hat{\eta} \mid Z^n} \mid Z^{i-1}} [D^*(q \| p_{\breve{W}_\theta})] \right] \leq \\ &\tfrac{C_{2\eta}}{n} \mathbf{E}^*_{Z^n} \left[\textstyle\sum_{i=1}^n \mathbf{E}_{\theta \sim \breve{W}_{\hat{\eta} \mid Z^n} \mid Z^{i-1}} \left[- \ln \frac{p_\theta(Z_i)}{q(Z_i)} \right] + \frac{u - \ln \pi(\hat{\eta} \mid Z^n)}{\eta} \right] \end{aligned} \tag{21}$$

The left side is equal to $\mathbf{E}^*_{Z^n} \left[D^*(q \| p_{\text{CES}(\breve{W}_{\hat{\eta} \mid Z^n}; Z^n)}) \right]$. We now take π to be the uniform prior on $\mathcal{S}_n$, so that for all $\eta \in \mathcal{S}_n$, $- \ln \pi(\eta) = \ln \|\mathcal{S}_n\| = \ln \|\kappa_{\max} + 1\| = O(\ln \ln n)$. The result now follows from (21), noting that the right-hand side increases if we replace $\hat{\eta} \mid Z^n$ by $\eta \in \mathcal{S}_n$.

Towards Theorem 3 We extend an inequality which, in various guises and level of detail, was proven earlier by M. Seeger (2002), D. McAllester (2003), O. Catoni (2007), J.Y. Audibert (2004) (in the context of PAC-Bayesian inference; see Zhang (2006b) for references to additional relevant papers by these authors), and A. Barron (with T. Cover (1991) and with J. Li (1999)), and T. Zhang (2006a,b) in the context of MDL-type inference. Our version is a direct extension of Theorem 2.1. of Zhang (2006b). Let $Z_1, Z_2, \ldots$, P^*, f_ν and $\mathcal{G}$ be as above Lemma 1, except that now we do *not* require $Z_1, Z_2, \ldots$ to be i.i.d. All earlier guises of Zhang's result assumed that $Z_1, \ldots, Z_n$ are i.i.d. both according to the 'true' P^* and according to all 'densities' $f_\nu(Z_i \mid z^{i-1})$, which were not allowed to depend on z^{i-1}. Our application of the inequality to prove Theorem 2 requires us to extend it to non-i.i.d. models (represented below by $f_\nu(Z_i \mid z^{i-1})$ which vary with z^{i-1}). As a by-product, we also extend it to non-i.i.d. Z_i (in principle this should allow us to extend the in this paper to some non-i.i.d. misspecification settings as considered by Shalizi (2009)). The result compares the expectation of $Z_i \mid Z^{i-1}$ to its actually observed value, and then takes another expectation over the values that can actually be observed. To ease readability, we denote the Z_i in the inner expectation as $\bar{Z}_i$.

Theorem 3. [Extended Zhang's Inequality] *For arbitrary $\mathcal{G}$, let Π (a "prior") be a distribution on $\mathcal{G}$ and let $\breve{W} : \bigcup_{n \geq 0} \mathcal{Z}^n \rightarrow \mathcal{G}$ be a randomized estimator. Then, with $D(\cdot \| \cdot)$ denoting KL divergence, we have:*

$$\begin{aligned} &\mathbf{E}^*_{Z^n} \mathbf{E}_{\nu \sim \breve{W} \mid Z^n} \left[\textstyle\sum_{i=1}^n - \ln \mathbf{E}^*_{\bar{Z}_i \mid Z^{i-1}} [f_\nu(\bar{Z}_i \mid Z^{i-1})] \right] \leq \\ &\quad \mathbf{E}^*_{Z^n} \left[\mathbf{E}_{\nu \sim \breve{W} \mid Z^n} \left[\textstyle\sum_{i=1}^n - \ln f_\nu(Z_i \mid Z^{i-1}) \right] + D(\, (\breve{W} \mid Z^n) \, \| \Pi) \right]. \end{aligned} \tag{22}$$

As a special case, suppose $\mathcal{G}$ is countable, π is a probability mass function on $\mathcal{G}$, and $\breve{\nu}$ is a deterministic estimator. Then

$$\begin{aligned} &\mathbf{E}^*_{Z^n} \left[\textstyle\sum_{i=1}^n - \ln \mathbf{E}^*_{\bar{Z}_i \mid Z^{i-1}} [f_{\breve{\nu} \mid Z^n}(\bar{Z}_i \mid Z^{i-1})] \right] \leq \\ &\mathbf{E}^*_{Z^n} \left[\textstyle\sum_{i=1}^n [- \ln f_{\breve{\nu} \mid Z^n}(Z_i \mid Z^{i-1})] - \ln \pi(\breve{\nu}) \right]. \end{aligned} \tag{23}$$

Theorem 2.1. of Zhang (2006b) is the special case for functions f_ν that satisfy, for all $i < n, z^i \in \mathcal{Z}^i$, $f_\nu(Z_i \mid z^{i-1}) = g_\nu(Z_i)$ for some fixed g_ν. The proof is in Appendix A.

Bibliography

J.Y. Audibert. *PAC-Bayesian statistical learning theory.* PhD thesis, Université Paris VI, 2004.

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

P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

O. Catoni. *PAC-Bayesian Supervised Classification.* Lecture Notes-Monograph Series. IMS, 2007.

Kamalika Chaudhuri, Yoav Freund, and Daniel Hsu. A new parameter-free hedging algorithm. In *NIPS-09*, 2009.

A.P. Dawid. Present position and potential developments: Some personal views, statistical theory, the prequential approach. *Journal of the Royal Statistical Society, Series A*, 147(2):278–292, 1984.

J.L. Doob. Application of the theory of martingales. In *Le Calcul de Probabilités et ses Applications. Colloques Internationaux du Centre National de la Recherche Scientifique*, pages 23–27, Paris, 1949.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

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

P. Grünwald. Safe learning: bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity. In *Proceedings of the Twenty-Fourth Conference on Learning Theory (COLT' 11)*, 2011.

P. Grünwald and J. Langford. Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, 66(2-3):119–149, 2007. DOI 10.1007/s10994-007-0716-7.

B. Kleijn and A. van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics*, 34(2), 2006.

S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:76–86, 1951.

J.Q. Li. *Estimation of Mixture Models.* PhD thesis, Yale University, New Haven, CT, 1999.

E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *Annals of Statistics*, 27:1808–1829, 1999.

D. McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51 (1):5–21, 2003.

M. Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, 2002.

C. Shalizi. Dynamics of bayesian updating with dependent data and misspecified models. *Electronic Journal of Statistics*, 3:1039–1074, 2009.

J. Takeuchi and A. R. Barron. Robustly minimax codes for universal data compression. In *Proceedings of the Twenty-First Symposium on Information Theory and Its Applications (SITA '98)*, Gifu, Japan, 1998.

A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32:135–166, 2004.

A. van der Vaart. *Asymptotic Statistics.* Cambridge University Press, Cambridge, UK, 1998.

T. van Erven, P. Grünwald, W. Koolen, and S. de Rooij. Adaptive hedge. In *Advances in Neural Information Processing Systems 24 (NIPS-11)*, 2011.

V. Vovk. Competitive on-line statistics. *Intern. Stat. Rev.*, 69:213–248, 2001.

V.G. Vovk. Aggregating strategies. In *Proc. COLT' 90*, pages 371–383, 1990.

Ziheng Yang. Fair-balance paradox, star-tree paradox, and bayesian phylogenetics. *Journal of Molecular Biology and Evolution*, 24(8):1639–1655, 2007.

Tong Zhang. From ϵ-entropy to KL entropy: analysis of minimum information complexity density estimation. *Annals of Statistics*, 34(5):2180–2210, 2006a.

Tong Zhang. Information theoretical upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006b.

A Additional Proofs

Existence of Best-Approximating Density (Section 3)

For given $\mathcal{P}$ and P^*, fix an arbitrary $p' \in \mathcal{P}$. There is a 1-to-1 correspondence between $\mathcal{P}$ and the set of random variables $\mathcal{L} := \{L_p : p \in \mathcal{P}\}$, where $L_p := -\log p(Z)/p'(Z)$, $Z \sim P^*$. By our assumption (3) that $U = U(\mathcal{P}) < \infty$, these random variables are uniformly bounded by $-\log U$ and $\log U$, so by Prohorov's theorem (van der Vaart, 1998), the set $\mathcal{L}$ is relatively compact in the topology of weak convergence. This implies that there exists a function $\tilde{L} : \mathcal{Z} \rightarrow [-\log U, \log U]$ such that $\mathbf{E}^*[\tilde{L}] = \inf_{p \in \mathcal{P}} D^*(p' \| p)$, where we used the notation of (5). By our assumption that the ratio in (3) is always well-defined, we have that $p'(z) > 0$ for all $z \in \mathcal{Z}$. Hence there exists a function $q : \mathcal{Z} \rightarrow \mathbb{R}^+$ such that for all $z \in \mathcal{Z}$, $\tilde{L}(z) = -\log q(z) + \log p'(z)$. For this q we have (a) $D(P^* \| q) = \inf_{p \in \mathcal{P}} D(P^* \| p)$. By definition of $\tilde{L}$ we must have $\sup_{z \in \mathcal{Z}} \{q(z)/p'(z), p'(z)/q(z)\} \leq U$. Since, for any $p \in \mathcal{P}$ we can write $\tilde{L}(z) = -\log q(z)/p(z) + \log p(z)/p'(z)$ we must also have, (b), $U(\mathcal{P} \cup \{q\}) \leq 2U(\mathcal{P}) = V(\mathcal{P})$. Together, (a) and (b) show that q is a best-approximating density.

We note that we did not prove that $\int q d\mu < \infty$; it is thus conceivable that $\int q d\mu = \infty$ such that q cannot be used to make probabilistic predictions. But this is of no consequence, since q is only ever used in this paper as a 'gold standard' to which other densities p are compared. We never use q itself, neither to make a prediction nor as a part of a mixture with other p.

Proof of Theorem 3

Without loss of generality, we may assume that $\mathit{\Pi}$ has density π and, for each i, z^i, $\breve{W} \mid z^i$ has density $\breve{w}(\cdot \mid z^i)$, where all these densities are with respect to a common underlying measure. Define, for all ν, the function $h_\nu : \bigcup_{m=0}^{n-1} \mathcal{Z}^m \to \mathbb{R}^+$ by

$$h_\nu(z^m) := \mathbf{E}^*_{Z_{m+1},\ldots,Z_n \mid z^m} \left[\frac{\prod_{i=m+1}^n f_\nu(Z_i \mid z^m, Z_{m+1}, \ldots, Z_{i-1})}{\prod_{i=m+1}^n \mathbf{E}^*_{\bar{Z}_i \mid Z^{i-1}}[f_\nu(\bar{Z}_i \mid z^m, Z_{m+1}, \ldots, Z_{i-1})]} \right],$$

with $h_\nu(z^m) = 0$ if the denominator is 0.

Lemma 3. *For all $\nu \in \mathcal{G}$, all $m < n$, $P^*(h_\nu(Z^m) = 1) = 1$.*

Proof. For $m = n - 1$, $h_\nu(Z^m)$ evaluates to 1 P^*-almost surely and the result is obvious. We now show that if the result holds for any $0 < m < n$, it also holds for $m - 1$, thus proving the result by induction. For convenience we also write Z_j^i to denote $Z_j, Z_{j+1}, \ldots, Z_i$.

Suppose then that $h_\nu(Z^m) = 1$, P^*-a.s. for some $m \geq 1$. We can write

$$h_\nu(Z^{m-1}) = \mathbf{E}^*_{Z_m^n \mid Z^{m-1}} \left[\frac{\prod_{i=m}^n f_\nu(Z_i \mid Z^{i-1})}{\prod_{i=m}^n \mathbf{E}^*_{\bar{Z}_i \mid Z^{i-1}}[f_\nu(\bar{Z}_i \mid Z^{i-1})]} \right] =$$

$$\underset{Z_m \mid Z^{m-1}}{\mathbf{E}^*} \; \underset{Z_{m+1}^n \sim P^* \mid Z^{m-1}, Z_m}{\mathbf{E}^*} \left[\frac{f_\nu(Z_m \mid Z^{m-1}) \cdot \prod_{i=m+1}^{n} f_\nu(Z_i \mid Z^{i-1})}{\mathbf{E}^*_{\bar{Z}_m \mid Z^{m-1}}[f_\nu(\bar{Z}_m \mid Z^{m-1}) \cdot \prod_{i=m+1}^{n} \mathbf{E}^*_{\bar{Z}_i \mid Z^{i-1}}[f_\nu(\bar{Z}_i \mid Z^{i-1})]} \right] =$$

$$\underset{Z_m \mid Z^{m-1}}{\mathbf{E}^*} \left[\frac{f_\nu(Z_m \mid Z^{m-1})}{\mathbf{E}^*_{\bar{Z}_m \mid Z^{m-1}}[f_\nu(\bar{Z}_m \mid Z^{m-1})]} \cdot \underset{Z_{m+1}^n \sim P^* \mid Z^{m-1}, Z_m}{\mathbf{E}^*} \left[\frac{\prod_{i=m+1}^{n} f_\nu(Z_i \mid Z^{i-1})}{\prod_{i=m+1}^{n} \mathbf{E}^*_{\bar{Z}_i \mid Z^{i-1}}[f_\nu(\bar{Z}_i \mid Z^{i-1})]} \right] \right] =$$

$$\mathbf{E}^*_{Z_m \mid Z^{m-1}} \left[\frac{f_\nu(Z_m \mid Z^{m-1})}{\mathbf{E}^*_{\bar{Z}_m \mid Z^{m-1}}[f_\nu(\bar{Z}_m \mid Z^{m-1})]} \cdot h(Z^m) \right] \overset{(a)}{=}$$

$$\mathbf{E}^*_{Z_m \mid Z^{m-1}} \left[\frac{f_\nu(Z_m \mid Z^{m-1})}{\mathbf{E}^*_{\bar{Z}_m \mid Z^{m-1}}[f_\nu(\bar{Z}_m \mid Z^{m-1})]} \cdot 1 \right] \overset{(b)}{=} 1,$$

where (a) follows by induction and (b) is immediate.

In particular, note that $h_\nu(Z^0)$ (a number rather than a function) is well-defined, and the lemma implies that $h_\nu(Z^0) = 1$, P^*-almost surely, so that if we set

$$S_{n,\nu} := \frac{\prod_{i=1}^n f_\nu(Z_i \mid Z^{i-1})}{\prod_{i=1}^n \mathbf{E}^*_{\bar{Z}_i \mid Z^{i-1}}[f_\nu(\bar{Z}_i \mid Z^{i-1})]}$$

we have for all $\nu \in \mathcal{G}$ $\mathbf{E}^*_{Z^n}[S_{n,\nu}] = h_\nu(Z^0) = 1$ and hence $\mathbf{E}^*_{Z^n}[\mathbf{E}_{\nu \sim \mathit{\Pi}}[S_{n,\nu}]] = 1$ for any prior $\mathit{\Pi}$, in particular the one we chose. The result follows immediately upon noting (with D abbreviating $D(\breve{W}|Z^n \,\|\varPi))$

$$\begin{aligned}
&\underset{Z^n}{\mathbf{E}^*} \underset{\nu\sim\breve{W}|Z^n}{\mathbf{E}} \left[\sum_{i=1}^n - \ln \underset{\bar{Z}_i|Z^{i-1}}{\mathbf{E}^*} [f_\nu(\bar{Z}_i \mid Z^{i-1})] - \left[\sum_{i=1}^n [- \ln f_\nu(Z_i \mid Z^{i-1})] + D\right]\right] = \\
&\underset{Z^n}{\mathbf{E}^*} \underset{\nu\sim\breve{W}|Z^n}{\mathbf{E}} \left[- \ln \tfrac{\breve{w}(\nu|Z^n)}{\pi(\nu)} - \sum_{i=1}^n \ln \underset{\bar{Z}_i|Z^{i-1}}{\mathbf{E}^*} [f_\nu(\bar{Z}_i \mid Z^{i-1})] + \sum_{i=1}^n \ln f_\nu(Z_i \mid Z^{i-1})\right] = \\
&\mathbf{E}^*_{Z^n} \mathbf{E}_{\nu\sim\breve{W}|Z^n} \left[\ln \left(\tfrac{\pi(\nu)}{\breve{w}(\nu|Z^n)} \cdot \tfrac{\prod_{i=1}^n f_\nu(Z_i|Z^{i-1})}{\prod_{i=1}^n \mathbf{E}^*_{\bar{Z}_i|Z^{i-1}}[f_\nu(\bar{Z}_i|Z^{i-1})]}\right)\right] \leq \\
&\mathbf{E}^*_{Z^n} \left[\ln \mathbf{E}_{\nu\sim\breve{W}|Z^n} \left[\tfrac{\pi(\nu)}{\breve{w}(\nu|Z^n)} \cdot \tfrac{\prod_{i=1}^n f_\nu(Z_i|Z^{i-1})}{\prod_{i=1}^n \mathbf{E}^*_{\bar{Z}_i|Z^{i-1}}[f_\nu(\bar{Z}_i|Z^{i-1})]}\right]\right] = \\
&\mathbf{E}^*_{Z^n} [\ln \mathbf{E}_{\nu\sim\varPi}[S_{n,\nu}]] \leq \ln \mathbf{E}^*_{Z^n} \mathbf{E}_{\nu\sim\varPi}[S_{n,\nu}] = \ln 1 = 0,
\end{aligned}$$

where the equalities are just rearranging and the inequalities are both Jensen's. This proves (22). To show that (23) is a special case, suppose that $\mathcal{G}$ is countable, $\varPi$ has mass function π and $\breve{W}$ represents the deterministic estimator $\breve{\nu}$, i.e. for all $z^n \in \mathcal{Z}^n$, $\breve{W} \mid z^n$ has mass function $\breve{w}(\nu \mid z^n)$ with $\breve{w}((\breve{\nu}| \, z^n) \mid z^n) = 1$. Then all expectations of $f_\nu(\cdot \mid \cdot)$ over $\nu \sim \breve{W} \mid Z^n$ can be replaced by $f_{\hat{\nu}}(\cdot \mid \cdot)$. (23) now follows because

$$D(\, (\breve{W}|z^n) \, \|\varPi) = \sum_\nu \breve{w}(\nu \mid z^n) \log \frac{\breve{w}(\nu \mid z^n)}{\pi(\nu)} = - \log \pi(\, (\breve{\nu} \mid z^n) \,).$$

The result is proven.