TCS-TR-A-06-11

TCS Technical Report

Potential Functions for Stochastic Model Selection

by

J. Poland

Division of Computer Science

Report Series A

March 3, 2006

Hokkaido University
Graduate School of
Information Science and Technology

Email: jan@ist.hokudai.ac.jp

Phone: +81-011-706-7675
Fax: +81-011-706-7675

TCS-TR-A-06-11 March 3, 2006

Potential Functions for Stochastic Model Selection

Jan Poland*
Graduate School of Information Science and Technology
Hokkaido University, Japan
jan@ist.hokudai.ac.jp
http://www-alg.ist.hokudai.ac.jp/~jan

Abstract

We prove performance guarantees for Bayesian learning algorithms, in particular stochastic model selection, with the help of potential functions. Such a potential quantifies the current state of learning in the system, in a way that the expected error in the next step is bounded by the expected decrease of the potential. For Bayesian stochastic model selection, an appropriate potential function will be specified by introducing the *entropy potential*, a quantity which we define as the worst-case entropy of a model class with regard to the true model. The resulting cumulative error bounds correspond to Solomonoff's theorem and are essentially sharp. They imply consistency, namely almost sure convergence of the predictive probabilities to the true ones, and loss/regret bounds for arbitrary bounded loss function. Although we formulate our results in the classification framework, they are equally applicable to the prediction of non-i.i.d. sequences.

1 Introduction

Most of the performance guarantees proven recently in learning theory fall into the category of "data dependent bounds". They are powerful for constructing learning algorithms and usually do not need any assumption on the data generating process. On the other hand, if we want to make assertions about a learner's (expected) quality, before we have seen any data, a different type of performance guarantees is required: bounds that hold prior to any observation under certain assumptions on the data generating process. This is the topic of the present paper.

An archetype for this latter sort of performance guarantees is *Solomonoff's universal induction result* [Sol78] (see Theorem 3 below). It gives a tight bound on the

*This work was supported by JSPS 21st century COE program C01.

expected cumulative quadratic error of a Bayesian mixture learner, provided that the learner is based on a countable model class containing the data generating distribution. This does not only imply a very strong consistency assertion (namely almost sure convergence, i.e. convergence with probability one), but also loss bounds for arbitrary loss function. A variant for continuously parameterized model class has been given in [CB90]. *Minimum description length* (MDL) defines another class of learners for which corresponding results are known [Ris96, PH05].

While, for given data, MDL and Bayes mixture learners make deterministic predictions, this is different for *stochastic model selection*, sometimes also referred to as Gibbs sampling. For this learner we know of no corresponding theorem so far: This is the main focus of the present paper. A different important type of theorems has been intensely studied for stochastic model selection in the recent past, namely *PAC-Bayesian theorems*, e.g. [McA03]: They state that PAC bounds for a single model carry over to a learner which selects models randomly according to a suitable posterior distribution.

By proving cumulative error bounds for the stochastic model selection learner, we will obtain two conclusions: We will show that this learner is *consistent* in the (strong) sense that the predictive probabilities converge to the true ones almost surely. Moreover we will obtain loss or regret bounds under arbitrary bounded loss functions. Both results are, to our knowledge, new for this learner.

The present proofs are based on the method of *potential functions*. A potential quantifies the current state of learning, such that the expected error in the next step does not exceed the expected decrease of the potential function in the next step. If we then can bound the cumulative decrease of the potential function, we obtain the desired bounds. Again Solomonoff's result can be understood as an archetype for this method, as we will see below. The potential method used here has been inspired by a related (but technically different) proof technique in prediction with expert advice [CBL03]. Also, note the fundamental difference between Bayesian stochastic model selection, which we focus on, and the very popular stochastic expert selection algorithms, e.g. [FS97]. In the latter case, the posterior, according to which is sampled, is not Bayesian, but specifically designed for regret minimization. Expert algorithms do not yield estimates for the probabilities governing the data.

In order to get a potential function for stochastic model selection satisfying the desired properties, we will introduce the *entropy potential*. This quantity is defined as the worst-case entropy of the model class under all admissible transformations of the weights, where the weight of the true distribution (which is in the model class by assumption) is kept fixed. The entropy potential is possibly a novel definition in this work.

This paper is structured as follows. In the next section, we will introduce the notation and prove Solomonoff's result with a potential function. In Section 3, we consider stochastic model selection and prove the main auxiliary result in order to obtain bounds. Section 4 defines the entropy potential and proves bounds for general

countable model class. In Section 5 we turn to the question how large the newly defined entropy potential can be. The last section contains discussion and conclusions.

2 Setup and Bayes Mixture

We work in a general discrete Bayesian online classification framework with stochastic concepts. All our theorems and proofs carry over to the prediction of non-i.i.d. sequences (this setup is defined e.g. in [PH05], compare also Remark 2).

Let $\mathcal{X} = \{1 \dots |\mathcal{X}|\}$ be a finite alphabet, $\mathcal{Z}$ be a finite or countable set of possible inputs (see again Remark 2), and $\mathcal{C} = \nu_1, \nu_2, \dots$ be a finite or countable model class. Each model $\nu \in \mathcal{C}$ specifies probability distributions[1] on $\mathcal{X}$ for all inputs $z \in \mathcal{Z}$, i.e. ν is a function

$$\nu : z \mapsto \big(\nu(x|z)\big)_{x\in\mathcal{X}} \text{ where } \nu(x|z) \geq 0 \text{ and } \sum_{x\in\mathcal{X}} \nu(x|z) = 1. \tag{1}$$

Each $\nu \in \mathcal{C}$ is assigned a prior weight $w_\nu > 0$, where $\sum_{\nu\in\mathcal{C}} w_\nu = 1$. (We need not consider models with zero prior weight, as they don't have any impact for anything of what follows.) In order to make clear that we talk of the prior or *initial* weight, opposed to a posterior weight, we will sometimes write w_ν^{init} instead of w_ν.

We assume that there is one data generating or *true* distribution $\mu \in \mathcal{C}$. Then the online classification proceeds in discrete time $t = 1, 2, \dots$: An input z_t is generated by some mechanism (compare Remark 2 below) we must issue (i.e. the learner must compute) a guess $\big(p(x)\big)_{x\in\mathcal{X}}$ (where $\sum_{x\in\mathcal{X}} p(x) = 1$) for the current probability vector $\big(\mu(x|z_t)\big)_{x\in\mathcal{X}}$, and an outcome $x_t \in \mathcal{X}$ is sampled according to $\big(\mu(x|z_t)\big)$ and revealed to the learner (note that the probabilities $\big(\mu(x|z_t)\big)$ are *not* revealed).

After each observation x_t, we may update the weights w_ν by Bayes' rule, thus obtaining, after time $t-1$ and before time t, the posterior weights

$$w_\nu(h_{<t}) = w_\nu(h_{1:t-1}) = w_\nu(z_{1:t-1}, x_{1:t-1}) = \frac{w_\nu \prod_{i=1}^{t-1} \nu(x_i|u_i)}{\sum_{\nu'\in\mathcal{C}} w_{\nu'} \prod_{i=1}^{t-1} \nu'(x_i|u_i)},$$

where $h_{<t} = (z_{<t}, x_{<t}) = (u_1, x_1, u_2, x_2, \dots, z_{t-1}, x_{t-1})$ denotes the history. Then, in the Bayesian sense it is optimal to estimate the current probabilities according to the *Bayes mixture*

$$\xi(x|z_t, h_{<t}) = \sum_{\nu\in\mathcal{C}} w_\nu(h_{<t})\nu(x|z_t).$$

Example 1 Assume that $\mathcal{X}$ is binary and $\mathcal{Z}$ contains only a single element. In this case the observations are *Bernoulli* trials, i.e. they result from fair or unfair coin flips.

[1] We don't consider *semimeasures*, as our methods below rely on normalized probability distributions. This restriction can be possibly lifted to some extent, however we do not expect the consequences to be very interesting (see also Example 22).

$\mathcal{C}$ specifies the set of possible coins we consider, and it is well-known that all posterior weights but the weight of the true coin will converge to zero almost surely for $t \to \infty$. With the set of coins $\mathcal{C} \cong \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$ and the true coin being the fair one, it is easy to see that this example gives a lower bound $o(-\log w_\mu)$ on the expected quadratic error of Bayes mixture and stochastic model selection predictions, namely the l.h.s. expressions of (4) and (7), respectively.

Remark 2 The inputs z_t are not used at all throughout this paper, so the mechanism which generates them doesn't need to be specified. We could as well work in an input-less *sequence prediction* setup, which is common for Solomonoff induction (Theorem 3 below). We decided to keep the inputs, as stochastic model selection is usually considered in a classification setup (compare also Example 22). We incorporate the inputs into the history $h_{<t}$, thus they don't complicate the notation.

Solomonoff's [Sol78] remarkable universal induction result tightly bounds the performance guarantee for the Bayes mixture with an arbitrary input sequence z_t. For introductory purpose, we prove it here in the classification setup. We use an appropriate *potential function*, thereby slightly modifying the proof from [Hut04].

Theorem 3 (Solomonoff's universal induction result) *Assume that the data generating distribution is contained in the model class, i.e. $\mu \in \mathcal{C}$. Define the* complexity potential *as*

$$\mathcal{K}(h_{<t}) = -\log w_\mu(h_{<t}). \tag{2}$$

(All logarithms are natural in this paper.) For any current input z_t and any history $h_{<t}$, this potential satisfies

$$\begin{aligned} &(i) \quad \mathcal{K}(h_{<t}) \geq 0, \\ &(ii) \quad \mathcal{K}(h_{<t}) - \mathbf{E}_{x_t \sim \mu(\cdot|z_t)} \mathcal{K}(h_{1:t}) \geq \sum_{x \in \mathcal{X}} \big(\mu(x|z_t) - \xi(x|z_t, h_{<t})\big)^2. \end{aligned} \tag{3}$$

By summing up the expectation of (ii) while observing (i), we immediately obtain Solomonoff's assertion for arbitrary sequence of inputs $z_1, z_2, \ldots$:

$$\sum_{t=1}^{\infty} \mathbf{E} \|\mu - \xi\|_2^2 := \sum_{t=1}^{\infty} \mathbf{E} \big\|\mu(\cdot|z_t) - \xi(\cdot|z_t, h_{<t})\big\|_2^2 \leq \mathcal{K}^{\text{init}} = -\log w_\mu^{\text{init}}, \tag{4}$$

where expectation is with respect to μ, and the squared 2-norm of a vector $v \in \mathbb{R}^{|\mathcal{X}|}$ is defined as usual, $\|v\|_2^2 = \sum_i v_i^2$. As we will see in the proof of Theorem 7, this implies that the Bayes mixture probabilities ξ converge to the true probabilities μ almost surely.

Proof. Clearly, (i) holds. In order to show (ii), we observe that $w_\mu(h_{1:t}) = w_\mu(h_{<t}) \frac{\mu(x_t|z_t)}{\xi(x_t|z_t, h_{<t})}$. Then, simplifying the notation by suppressing the history $h_{<t}$ and the current input z_t (e.g. $\mathcal{K}$ stands for $\mathcal{K}(h_{<t})$),

$$\mathcal{K} - \mathbf{E}\mathcal{K}(x) = \mathcal{K} - \sum_{x \in \mathcal{X}} \mu(x) \left(\mathcal{K} - \log \tfrac{\mu(x)}{\xi(x)}\right) = D\big[\mu(\cdot|z_t) \big\| \xi(\cdot|z_t, h_{<t})\big].$$

The r.h.s. here is called *Kullback-Leibler divergence*. By the following lemma it is an upper bound for $\sum_{x\in\mathcal{X}}\big(\mu(x|z_t)-\xi(x|z_t,h_{<t})\big)^2$. □

Lemma 4 *For two probability distributions μ and ρ on $\mathcal{X}$, we have*

$$\sum_{a\in\mathcal{X}}\big(\mu(a)-\rho(a)\big)^2 \le \sum_{a\in\mathcal{X}}\mu(a)\log\tfrac{\mu(a)}{\rho(a)}.$$

This well known inequality is proven for instance in [Hut04, Sec.3.9.2].

By Kraft's inequality, the complexity $\mathcal{K}$ of μ can be interpreted as μ's description length. Thus, Solomonoff's theorem asserts that the predictive complexity (measured in terms of the quadratic error) coincides with the descriptive complexity, if the data is rich enough to distinguish the models. Then $\mathcal{K}$ can be viewed as the *state of learning* in the discrete model class. Observe that only the *expected* progress, i.e. decrease of $\mathcal{K}$, is positive. The actual progress depends on the outcome of x_t and is positive if and only if $\mu(x_t)\ge\xi(x_t)$. If the probability vectors μ and ξ coincide, then – according to this potential function – no learning takes place for any observation, as then $\mathcal{K}(x_t)=\mathcal{K}$ for all x_t. Hence, the complexity potential $\mathcal{K}$ need not always be a good choice to describe the learning state.

Example 5 Consider a binary alphabet and a model class containing three distributions ν_1,ν_2,ν_3, predicting $\nu_i(1|z)=\frac{i}{4}$ for some input z. Suppose $\mu=\nu_2$, i.e. the true probability is $\frac{1}{2}$. Then we cannot measure the learning progress after the observation in terms of $\mathcal{K}$. However, there should be a progress, and indeed there is one, if we consider the *entropy* of the model class. This will become clear with Lemma 6.

3 Stochastic Model Selection

Here is another case where the complexity potential $\mathcal{K}$ is not appropriate to quantify the state of learning. In *stochastic model selection*, the current prediction vector $\Xi(\cdot|z_t,h_{<t})$ is obtained by randomly sampling a model according to the current weights $w_\nu(h_{<t})$ and using this model's prediction, i.e.

$$\Xi(\cdot|z_t,h_{<t})=\nu_J(\cdot|z_t) \text{ where } \mathbf{P}(J=i)=w_{\nu_i}(h_{<t}).$$

Hence, Ξ is a random variable depending on the sampled index J. The following lemma gives a first indication for a suitable potential function for learning with stochastic model selection.

Lemma 6 *Assume that the current entropy of the model class,*

$$\mathcal{H}(h_{<t})=-\sum_{\nu\in\mathcal{C}}w_\nu(h_{<t})\log w_\nu(h_{<t}),$$

is finite. Then, for any input z_t,

$$\mathcal{H}(h_{<t}) - \mathbf{E}_{x_t\sim\xi(\cdot|z_t,h_{<t})}\mathcal{H}(h_{1:t}) = \sum_{\nu\in\mathcal{C}} w_\nu(h_{<t}) \sum_{x\in\mathcal{X}} \nu(x|z_t) \log \tfrac{\nu(x|z_t)}{\xi(x|z_t,h_{<t})}$$
$$\geq \sum_{\nu\in\mathcal{C}} w_\nu(h_{<t}) \sum_{x\in\mathcal{X}} \big(\nu(x|z_t) - \xi(x|z_t,h_{<t})\big)^2 =: \mathbf{E}\big\|\Xi - \xi\big\|_2^2.$$

Proof. The equality is straightforward computation. Then use Lemma 4 for the inequality. □

Unfortunately, the l.h.s. of the above inequality contains an expectation w.r.t. ξ instead of μ. Since on the other hand μ governs the process and generally differs from ξ, the entropy $\mathcal{H}$ is not directly usable as a potential for the Ξ's deviation from its mean ξ. The following theorem demonstrates an easy fix, which however exponentially blows up the potential.

Theorem 7 (Predictive performance of stochastic model selection: loose bound) *Assume that $\mu \in \mathcal{C}$. Define the potential $\mathcal{P}_E(h_{<t}) = \mathcal{H}(h_{<t}) \exp\big(\mathcal{K}(h_{<t})\big) = \mathcal{H}(h_{<t})/w_\mu(h_{<t})$. Then, for any history $h_{<t}$ and any current input z_t,*

$$\mathcal{P}_E(h_{<t}) - \mathbf{E}_{x_t\sim\mu(\cdot|z_t)}\mathcal{P}_E(h_{1:t}) \geq \mathbf{E}\big\|\Xi(\cdot|z_t,h_{<t}) - \xi(\cdot|z_t,h_{<t})\big\|_2^2. \tag{5}$$

Consequently, with $\mathcal{H}^{\text{init}} = -\sum_{\nu\in\mathcal{C}} w_\nu^{\text{init}} \log w_\nu^{\text{init}}$ denoting the initial entropy,

$$\sum_{t=1}^{\infty} \mathbf{E}\big\|\Xi - \xi\big\|_2^2 \leq \mathcal{P}_E^{\text{init}} = \mathcal{H}^{\text{init}}/w_\mu^{\text{init}}, \tag{6}$$

$$\sum_{t=1}^{\infty} \mathbf{E}\big\|\Xi - \mu\big\|_2^2 \leq -\log(w_\mu^{\text{init}}) + \mathcal{H}^{\text{init}}/w_\mu^{\text{init}} + 2\sqrt{-\mathcal{H}^{\text{init}} \log(w_\mu^{\text{init}})/w_\mu^{\text{init}}}, \tag{7}$$

and the predictions by Ξ converge to the true probabilities μ almost surely.

Proof. Recall $w_\mu(h_{1:t}) = w_\mu(h_{<t}) \frac{\mu(x_t|z_t)}{\xi(x_t|z_t,h_{<t})}$. Since always $1/w_\mu(h_{<t}) \geq 1$, using Lemma 6 we obtain (5) by

$$\mathcal{P}_E(h_{<t}) - \sum_{x\in\mathcal{X}} \mu(x|z_t)\mathcal{P}_E(h_{1:t}) = \tfrac{1}{w_\mu(h_{<t})}\big(\mathcal{H}(h_{<t}) - \sum_{x\in\mathcal{X}} \xi(x|z_t,h_{<t})\mathcal{H}(h_{1:t})\big)$$
$$\geq \mathbf{E}\big\|\Xi(\cdot|z_t,h_{<t}) - \xi(\cdot|z_t,h_{<t})\big\|_2^2.$$

Summing the expectation up yields (6). Using this together with (4) and the triangle inequality $\sqrt{\sum \mathbf{E}\big\|\Xi - \mu\big\|_2^2} \leq \sqrt{\sum \mathbf{E}\big\|\Xi - \xi\big\|_2^2} + \sqrt{\sum \mathbf{E}\big\|\xi - \mu\big\|_2^2}$, we conclude (7). Finally, almost sure convergence follows from

$$\mathbf{P}\Big(\exists t \geq n : s_t \geq \varepsilon\Big) = \mathbf{P}\Big(\bigcup_{t\geq n} \big\{s_t \geq \varepsilon\big\}\Big) \leq \sum_{t\geq n} \mathbf{P}\big(s_t \geq \varepsilon\big) \leq \frac{1}{\varepsilon}\sum_{t=n}^{\infty} \mathbf{E}s_t \stackrel{n\to\infty}{\longrightarrow} 0$$

for each $\varepsilon > 0$, with $s_t = \mathbf{E}\big\|\Xi(\cdot|z_t, h_{<t}) - \mu(\cdot|z_t, h_{<t})\big\|_2^2$. □

In particular, this theorem shows that the entropy of a model class, if it is initially finite, necessarily remains finite almost surely. Moreover, it establishes almost sure asymptotic consistency of prediction by stochastic model selection in our Bayesian framework. However, it does not provide meaningful error bounds for all but very small model classes, since the r.h.s. of the bound is exponential in the complexity, hence possibly huge.

Before continuing to show better bounds, we demonstrate that the entropy is indeed a lower bound for any successful potential function for stochastic model selection.

Example 8 Let the alphabet be binary. Let $w_\mu = 1 - \frac{1}{n}$, in this way $\mathcal{K} \approx \frac{1}{n}$ and can be made arbitrary small for large $n \in \mathbb{N}$. Fix a target entropy $H_0 \in \mathbb{N}$ and set $K = 2^{nH_0}$. Choose a model class that consists of the true distribution, always predicting $\frac{1}{2}$, and K other distributions with the same prior weight $1/(nK)$. In this way, the entropy of the model class is indeed close to $H_0 \log 2$. Let the input set be $\mathcal{Z} = \{1 \dots nH_0\}$, and let $\nu_b(1|z) = b_z$, where b_z is the zth bit of ν's index b in binary representation. Then it is not hard to see that on the input stream $z_{1:nH_0} = 1, 2, \dots nH_0$ always $\mu = \xi$. Moreover, at each time, $E\|\Xi - \mu\|_2^2 = 1/(4n)$. Therefore the cumulative error is $H_0/4$, i.e. of order of the entropy. Note that this error, which can be chosen arbitrarily large, is achievable for arbitrarily small complexity $\mathcal{K}$.

In the proof of Theorem 7, we used only one "wasteful" inequality, namely $1/w_\mu(h_{<t}) \geq 1$. The following lemma will be our main tool for obtaining better bounds.

Lemma 9 (Predictive performance of stochastic model selection, main auxiliary result) *Suppose that we have some function $B(h_{<t})$, depending on the history, with the following properties:*

(i) $B(h_{<t}) \geq \mathcal{H}(h_{<t})$ *(dominates the entropy),*

(ii) $\mathbf{E}_{x_t \sim \mu(\cdot|z_t)} B(h_{1:t}) \leq B(h_{<t})$ *(decreases in expectation),*

(iii) *the value of $B(h_{<t})$ can be approximated arbitrarily close by restricting to a finite model class.*

Then, for any history and current input, the potential function defined by

$$\mathcal{P}(h_{<t}) = \big[\mathcal{K}(h_{<t}) + \log(1 + \mathcal{H}(h_{<t}))\big](1 + B(h_{<t}))$$

satisfies

$$\mathcal{P}(h_{<t}) - \mathbf{E}_{x_t \sim \mu(\cdot|z_t)} \mathcal{P}(h_{1:t}) \geq \mathcal{H}(h_{<t}) - \mathbf{E}_{x_t \sim \xi(\cdot|z_t, h_{<t})} \mathcal{H}(h_{1:t}). \tag{8}$$

Proof. Because of (iii), we need to prove the lemma only for finite model class, the countable case then follows by approximation. In this way we avoid dealing with a Lagrangian on an infinite dimensional space below.

Again we drop all dependencies on the history $h_{<t}$ and the current input z_t from the notation. Then observe that in the inequality chain

$$\begin{aligned}
&\mathcal{K} + \log(1+\mathcal{H}) - \sum_{x\in\mathcal{X}} \mu(x)\big[\mathcal{K}(x) + \log(1+\mathcal{H}(x))\big]\frac{1+B(x)}{1+B} \\
&\geq \mathcal{K} + \log(1+\mathcal{H}) - \sum_{x\in\mathcal{X}} \frac{\mu(x)(1+B(x))}{\sum_{x'}\mu(x')(1+B(x'))}\big[\mathcal{K}(x) + \log(1+\mathcal{H}(x))\big] \qquad (9)\\
&\geq \frac{\sum_\nu w_\nu \sum_x \nu(x)\log\frac{\nu(x)}{\xi(x)}}{1+B}, \qquad (10)
\end{aligned}$$

(9) follows from assumption (ii), so that we only need to show (10) in order to complete the proof. We will demonstrate an even stronger assertion:

$$\log(1+\mathcal{H}) - \sum_{x\in\mathcal{X}} \tilde{\mu}_x\big[\log(1+\mathcal{H}(x)) - \log\tfrac{\mu(x)}{\xi(x)}\big] \geq \frac{\sum_\nu w_\nu \sum_x \nu(x)\log\frac{\nu(x)}{\xi(x)}}{1+B} \qquad (11)$$

for any probability vector $\tilde{\mu} = (\tilde{\mu}_x)_{x\in\mathcal{X}} \in [0,1]^{|\mathcal{X}|}$ with $\sum_x \tilde{\mu}_x = 1$.

It is sufficient to prove (11) for all stationary points of the Lagrangian and all boundary points. In order to cover all of the boundary, we allow $\tilde{\mu}_x = 0$ for all x in some subset $\mathcal{X}_0 \subsetneq \mathcal{X}$ ($\mathcal{X}_0$ may be empty). Let $\tilde{\mathcal{X}} = \mathcal{X}\setminus\mathcal{X}_0$ and define $\xi(\tilde{\mathcal{X}}) = \sum_{x\in\tilde{\mathcal{X}}}\xi(x)$, $\xi(\mathcal{X}_0) = 1-\xi(\tilde{\mathcal{X}})$, and $\tilde{\xi}(x) = \xi(x)/\xi(\tilde{\mathcal{X}})$. Then (11) follows from

$$f(\tilde{\mu}) = \log(1+\mathcal{H}) - \sum_{x\in\tilde{\mathcal{X}}} \tilde{\mu}_x\big(\tilde{V}(x) - \log\tfrac{\mu(x)}{\tilde{\xi}(x)}\big) \geq \frac{\sum_\nu w_\nu \sum_x \nu(x)\log\frac{\nu(x)}{\xi(x)}}{1+B}, \qquad (12)$$

where $\tilde{V}(x) = \log(1 - \sum_\nu \frac{w_\nu\nu(x)}{\tilde{\xi}(x)}\log\frac{w_\nu\nu(x)}{\xi(x)})$.

We now identify the stationary points of the Lagrangian $\mathcal{L}(\tilde{\mu},\lambda) = f(\tilde{\mu}) - \lambda\big(\sum_x \tilde{\mu}_x - 1\big)$. The derivative of $\mathcal{L}$ w.r.t. all $\tilde{\mu}_x$ vanishes only if

$$\lambda = -\tilde{V}(x) + \log\tfrac{\mu(x)}{\tilde{\xi}(x)} \text{ for all } x\in\tilde{\mathcal{X}}. \qquad (13)$$

This implies $\mu(x) = \tilde{\xi}(x)e^{\lambda+\tilde{V}(x)}$, and, since the $\mu(x)$ sum up to one, $1 = e^\lambda\sum_x\tilde{\xi}(x)e^{\tilde{V}(x)}$. This can be reformulated as $\lambda = -\log\big[\sum_x\tilde{\xi}(x)e^{\tilde{V}(x)}\big]$. Using this and (13), (12) is rewrites as

$$\begin{aligned}
\frac{\sum_{\nu\in\mathcal{C}} w_\nu \sum_{x\in\mathcal{X}} \nu(x)\log\frac{\nu(x)}{\xi(x)}}{1+B} \quad &\leq \quad \log(1+\mathcal{H}) + \lambda \qquad (14)\\
&= \log(1 - \sum_{\nu\in\mathcal{C}} w_\nu\log w_\nu) - \log\big[1 - \sum_{x\in\tilde{\mathcal{X}}}\tilde{\xi}(x)\sum_{\nu\in\mathcal{C}}\tfrac{w_\nu\nu(x)}{\tilde{\xi}(x)}\log\tfrac{w_\nu\nu(x)}{\xi(x)}\big].
\end{aligned}$$

The arguments of both outer logarithms on the r.h.s. of (14) are at most $1+B$: For the left one this holds by assumption (i), $\mathcal{H}\leq B$, and for the right one also by (i) because $\mathbf{E}_{x\sim\xi}\mathcal{H}(x) \leq \mathcal{H}$. Since for $x\leq y\leq 1+B$ we have $\log(y)-\log(x) \geq \frac{y-x}{1+B}$, (14) follows from

$$\sum_{\nu\in\mathcal{C}} w_\nu \sum_{x\in\mathcal{X}_0}\nu(x)\log\tfrac{\nu(x)}{\xi(x)} \leq -\sum_{\nu\in\mathcal{C}} w_\nu \sum_{x\in\mathcal{X}_0}\nu(x)\log w_\nu.$$

But this relation is true by Jensen's inequality:

$$\sum_{\nu\in\mathcal{C}}\sum_{x\in\mathcal{X}_0}\tfrac{w_\nu\nu(x)}{\xi(\mathcal{X}_0)}\log\tfrac{w_\nu\nu(x)}{\xi(x)}\leq\log\Big(\sum_{\nu\in\mathcal{C}}\sum_{x\in\mathcal{X}_0}\tfrac{w_\nu\nu(x)}{\xi(\mathcal{X}_0)}\cdot\tfrac{w_\nu\nu(x)}{\xi(x)}\Big)\leq 0,$$

since the $\frac{w_\nu\nu(x)}{\xi(\mathcal{X}_0)}$ sum up to one and always $\frac{w_\nu\nu(x)}{\xi(x)}\leq 1$ holds. □

We now present a simple application of this result for finite model classes.

Theorem 10 (Predictive performance of stochastic model selection for finite model class) *Suppose that $\mathcal{C}$ consists of $N\in\mathbb{N}$ models, one of them is μ. Let*

$$\mathcal{P}_F(h_{<t})=\big[\mathcal{K}(h_{<t})+\log(1+\mathcal{H}(h_{<t}))\big](1+\log N).$$

Then $\mathcal{P}_F(h_{<t})-\mathbf{E}_{x_t\sim\mu}\mathcal{P}_F(h_{1:t})\geq\mathcal{H}(h_{<t})-\sum_{x\in\mathcal{X}}\xi(x|z_t,h_{<t})\mathcal{H}(h_{1:t})$ holds for any history $h_{<t}$ and current input z_t, Consequently,

$$\sum_{t=1}^{\infty}\mathbf{E}\big\|\Xi-\xi\big\|_2^2\leq\mathcal{P}_F^{\text{init}}=(1+\log N)\big[\log(1+\mathcal{H}^{\text{init}})-\log(w_\mu^{\text{init}})\big].\qquad(15)$$

Proof. Since the entropy of a class with N elements is at most $\log N$, this follows directly from Lemma 9. □

4 Entropy potential and countable classes

We now generalize Theorem 10 to arbitrary countable model classes. First note that there is one very convenient fact about the potential function proofs so far: (3), (5), and (8) all are *local* assertions, i.e. for a single time instance and history. If the local expected error is bounded by the expected potential decrease, then the desired consequence on the cumulative error holds.

The entropy cannot be directly used as B in Lemma 9, since it may increase under μ-expectation. Intuitively, the problem is the following: There could be a false model with a quite large weight, such that the entropy is kept "artificially" low. If this false model is now refuted with high probability by the next observation, then the entropy may (drastically) increase. An instance is constructed in the following example. Afterwards, we define the *entropy potential*, which does not suffer from this problem.

Example 11 Fix binary alphabet and let $\tilde{\mathcal{C}}$ and $\tilde{\mathcal{Z}}$ be model class and input space of Example 8. Let $\mathcal{C}=\tilde{\mathcal{C}}\cup\{\nu_{\text{fool}}\}$, $\mathcal{Z}=\tilde{\mathcal{Z}}\cup\{0\}$, $w_{\text{fool}}=1-\frac{1}{m}$, and the rest of the prior of mass $\frac{1}{m}$ be distributed to the other models as in Example 8. Also the true distribution remains the same one. If the input sequence is $z_{1:nH_0+1}=0,1,\ldots nH_0$, and $\nu_{\text{fool}}(1|0)=0$ while $\nu(1|0)=1$ for all other ν, then like before the cumulative error is (even more than) $H_0/4$, while the entropy can be made arbitrarily small for large m.

Definition 12 *(Entropy potential)* Let $H\big((w_\nu)_{\nu\in\mathcal{C}}\big) = -\sum_\nu w_\nu \log w_\nu$ be the entropy function. The *μ-entropy potential* (or short entropy potential) of a model class $\mathcal{C}$ containing the true distribution μ is

$$\Pi\big((w_\nu)_{\nu\in\mathcal{C}}\big) = \sup\big\{H\big((\tfrac{w_\nu p_\nu}{\sum_{\nu'} w_{\nu'} p_{\nu'}})_\nu\big) : p_\mu = 1 \wedge p_\nu \in [0,1]\ \forall \nu \in \mathcal{C}\setminus\{\mu\}\big\}. \qquad (16)$$

Here, the supremum is taken over all possible assignments of the *multiplying probabilities* $p_\nu \in [0,1]$, where $p_\mu = 1$.

Clearly, $\Pi \geq \mathcal{H}$. According to Theorem 7, Π is necessarily finite if $\mathcal{H}$ is finite, so the supremum can be replaced by a maximum. Note that the entropy potential is finitely approximable in the sense of (iii) in Lemma 9.

Proposition 13 (Characterization of Π) *For $S \subset \mathcal{C}$, let $w(S) = \sum_{\nu\in S} w_\nu$. There is exactly one subset $A \subset \mathcal{C}$ with $\mu \in A$, such that*

$$-\log w_\nu > L(A) := -\sum_{\nu'\in A} \tfrac{w_{\nu'}}{w(A)} \log w_{\nu'} \quad \Longleftrightarrow \quad \nu \in A\setminus\{\mu\}. \qquad (17)$$

We call A the set of active *models (in Π). Then, with $\log p_\nu = L(A) - \log w_\nu$ for $\nu \in \mathcal{C}\setminus S$, $\log p_\nu = 0$ for $\nu \in A\setminus\{\mu\}$, and $k = |\mathcal{C}\setminus A|$, we have*

$$\begin{aligned} \Pi \;=\; \Pi\big((w_\nu)_{\nu\in\mathcal{C}}\big) \;&=\; H\big((\tfrac{w_\nu p_\nu}{\sum_{\nu'} w_{\nu'} p_{\nu'}})_{\nu\in\mathcal{C}}\big) \\ &=\; \log\big(k + w(A) e^{L(A)}\big). \end{aligned} \qquad (18)$$

Moreover, this is scaling invariant in the weights, i.e. (17) yields the correct active set and (18) gives the correct value for weights that are not normalized, if these unnormalized weights are also used for computing $w(A)$ and $L(A)$.

Proof. We first argue that the maximum of (16) cannot be attained if some $p_\nu = 0$. To this aim, let $p \in [0,1]^{|\mathcal{C}|}$ be the multiplying probabilities, assume $p_\nu = 0$, and set $\tilde{H} = H\big((\tfrac{w_\nu p_\nu}{\sum_{\nu'} w_{\nu'} p_{\nu'}})_\nu\big)$. Now assume $p_\nu > 0$ and observe that

$$H\big((\tfrac{w_\nu p_\nu}{\sum_{\nu'} w_{\nu'} p_{\nu'}})_\nu\big) = -p_\nu w_\nu \log(p_\nu w_\nu) + (1 - p_\nu w_\nu)\big[-\log(1 - p_\nu w_\nu) + \tilde{H}\big] \geq \tilde{H}$$

holds if $-\log(p_\nu w_\nu) \geq \tilde{H}$. This can be realized for small enough $p_\nu > 0$, hence the maximum of (16) cannot be attained for $p_\nu = 0$.

Therefore, for a maximum of (16), we need that for each $\nu \in \mathcal{C}\setminus\{\mu\}$, either $p_\nu = 1$ or, with $w^p(\mathcal{C}) = \sum_\nu w_\nu p_\nu$ and $L^p(\mathcal{C}) = -\frac{1}{w^p(\mathcal{C})}\sum_\nu w_\nu p_\nu \log(w_\nu p_\nu)$,

$$0 = \frac{dH\big((\tfrac{w_\nu p_\nu}{\sum_{\nu'} w_{\nu'} p_{\nu'}})_\nu\big)}{dp_\nu} = \frac{w_\nu}{w^p(\mathcal{C})}\Big[-\log(w_\nu p_\nu) - L^p(\mathcal{C})\Big]. \qquad (19)$$

Those ν satisfying the latter condition have $\log p_\nu = -L(\mathcal{C}) - \log w_\nu$ and hence $L(\mathcal{C}) = L(\mathcal{C}\setminus\{\nu\})$. Therefore, each possible maximum of (16) corresponds to a subset $\tilde{A} \subset \mathcal{C}$

of *active models*, such that $\mu \in \tilde{A}$ and furthermore $p_\nu = 1$ for $\nu \in \tilde{A}$ and $\log p_\nu = -L(\tilde{A}) - \log w_\nu$ for $\nu \notin \tilde{A}$. Since only $\log p_\nu \leq 0$ is feasible, for $\nu \notin \tilde{A}$ we necessarily have $-\log w_\nu \leq L(\tilde{A})$. Subsets $\tilde{A}$ that satisfy this latter condition are called *feasible*.

Assume that we have a feasible subset $\tilde{A}$, then for all $\nu \notin \tilde{A}$, the complexity w_ν equals the average complexity of all $\nu \in \tilde{A}$. Hence

$$H\big((\tfrac{w_\nu p_\nu}{\sum_{\nu'} w_{\nu'} p_{\nu'}})_\nu\big) = -\sum_{\nu\in\tilde{A}} \frac{w_\nu}{w(\tilde{A})} \log \frac{w_\nu}{\sum_{\nu'\in\mathcal{C}} w_{\nu'}} = L(\tilde{A}) + \log\big(w(\tilde{A}) + ke^{-L(\tilde{A})}\big)$$
$$= \log\big(k + w(\tilde{A})e^{L(\tilde{A})}\big),$$

which proves (18) for any such $\tilde{A}$. Observe that our A defined in the assertion is the smallest feasible subset and therefore unique. So we only have to make sure no larger subset can result in a larger entropy.

To this aim, take any feasible subset $\tilde{A} \subset \mathcal{C}$. We assume that there is $\nu_1 \in \tilde{A} \setminus \{\mu\}$ such that $-\log w_{\nu_1} \leq L(\tilde{A})$. We need to show that then the entropy increases if we take out ν_1. But in this case, the derivative, computed as the r.h.s. of (19), is non-positive at $p_{\nu_1} = 1$. Thus we may increase the entropy by decreasing p_{ν_1} until the derivative vanishes. Repeating this step for all ν with the property $-\log w_\nu \leq L(\tilde{A})$, we conclude that the smallest feasible subset A gives the maximum entropy.

Finally, scaling invariance of the set (17) and the value (18) w.r.t. the weights is easy to see. □

The following result states that the the entropy potential *decreases in expectation*. This will allow us to obtain the desired bound with Lemma 9.

Theorem 14 *For any history $h_{<t}$ and current input z_t,*

$$\sum_{x_t\in\mathcal{X}} \mu(x_t|z_t)\Pi(h_{1:t}) \leq \Pi(h_{<t}).$$

Proof. We need to show the assertion only for finite model class: Once this is established, the general case follows by approximation.

Again, we drop the dependence on the history and the current input from the notation. We will show a slightly more general assertion: For any subset of the alphabet $\tilde{\mathcal{X}} \subset \mathcal{X}$, and any choice of probability vectors $\nu(x)$ for all $\nu \in \mathcal{C}$ we have

$$\sum_{x\in\tilde{\mathcal{X}}} \mu(x)\Pi(x) \leq \mu(\tilde{\mathcal{X}})\Pi\Big(\big[w_\nu \nu(\tilde{\mathcal{X}})\big]_{\nu\in\mathcal{C}}\Big), \tag{20}$$

where $\nu(\tilde{\mathcal{X}}) = \sum_{x\in\tilde{\mathcal{X}}} \nu(x)$ is the total ν-probability of the subset $\tilde{\mathcal{X}}$. We prove (20) by induction over the subset size $|\tilde{\mathcal{X}}|$. For $|\tilde{\mathcal{X}}| = 1$, there is nothing to show. If (20) holds for $\tilde{\mathcal{X}}$, then for $\tilde{\mathcal{X}}' = \tilde{\mathcal{X}} \cup \{x\}$,

$$\sum_{x\in\tilde{\mathcal{X}}'} \mu(x)\Pi(x) \leq \mu(\tilde{\mathcal{X}})\Pi\Big([w_\nu\nu(\tilde{\mathcal{X}})]_\nu\Big) + \mu(x)\Pi(x) \overset{(*)}{\leq} \mu(\tilde{\mathcal{X}}')\Pi\Big([w_\nu\nu(\tilde{\mathcal{X}}')]_\nu\Big)$$

implies the assertion. It remains to show $(*)$.

Now let $\tilde{w}_\nu = w_\nu \nu(\tilde{\mathcal{X}}')$ and $p_\nu = \nu(x)/\nu(\tilde{\mathcal{X}}')$ for all $\nu \in \mathcal{C}$, and set $\tilde{\mu} = p_\mu$. Then $(*)$ is equivalent to

$$(1-\tilde{\mu})\Pi\big([\tilde{w}_\nu(1-p_\nu)]_{\nu\in\mathcal{C}}\big) + \tilde{\mu}\Pi\big([\tilde{w}_\nu p_\nu]_{\nu\in\mathcal{C}}\big) \leq \Pi\big([\tilde{w}_\nu]_{\nu\in\mathcal{C}}\big), \tag{21}$$

where for all $\nu \in \mathcal{C}$ their values p_ν range in $p_\nu \in [0,1]$. Thus we have reduced the original assertion to binary alphabet.

In order to prove (21), it is sufficient to show that the maximum of the l.h.s. is attained if $p_\nu = \tilde{\mu}$ holds for all $\nu \in \mathcal{C}$. We first argue that the maximum can be only attained if in all three sets of weights, $[\tilde{w}_\nu]_\nu$, $[\tilde{w}_\nu(1-p_\nu)]_\nu$, and $[\tilde{w}_\nu p_\nu]_\nu$, the *same models are active* (see Proposition 13). Denote the respective sets of active models by A, A^0, A^1. Recall that the constructions in Proposition 13 do not require the weights to sum up to one, and define the quantities $\tilde{w}^1(A^1) = \sum_{\nu\in A^1}\tilde{w}_\nu p_\nu$ and $L^1(A^1) = -\sum_{\nu\in A^1}\frac{\tilde{w}_\nu p_\nu}{\tilde{w}^1(A^1)}\log(\tilde{w}_\nu p_\nu)$ and $\Pi^1 = \log\big(|\mathcal{C}\setminus A^1| + \tilde{w}^1(A^1)e^{L^1(A^1)}\big)$, and in the same way, the quantities $\tilde{w}^0(A^0)$, $L^0(A^0)$, and Π^0.

For active $\nu \in A^0$ or $\nu \in A^1$, respectively, the respective derivatives of Π^0 and Π^1 are computed as

$$\begin{aligned}
\frac{d\Pi^0}{dp_\nu} &= -\frac{\tilde{w}_\nu}{\Pi^0}e^{L^0(A^0)}\big(-\log[\tilde{w}_\nu(1-p_\nu)] - L^0(A^0)\big) < 0 \text{ for } \nu \in A^0\setminus\{\mu\} \text{ and}\\
\frac{d\Pi^1}{dp_\nu} &= \frac{\tilde{w}_\nu}{\Pi^1}e^{L^1(A^1)}\big(-\log[\tilde{w}_\nu p_\nu] - L^1(A^1)\big) > 0 \text{ for } \nu \in A^1\setminus\{\mu\},
\end{aligned}$$

where $\frac{d\Pi^1}{dp_\nu} > 0$ follows from $\big(-\log[\tilde{w}_\nu p_\nu] - L^1(A^1)\big) > 0$ for $\nu \in A^1$ (and analogously for Π^0). For inactive $\nu \notin A^0$ or $\nu \notin A^1$, respectively, the respective derivatives vanish.

Consider now a model $\nu \notin A$ which is inactive in Π. If we choose $p_\nu = \mu$, then it is inactive in both Π^0 and Π^1, i.e. both $\nu \notin A^0$ and $\nu \notin A^1$ hold. If we decrease p_ν until it becomes active in Π^1, then, because of $\frac{d\Pi^1}{dp_\nu} > 0$ and $\frac{d\Pi^0}{dp_\nu} = 0$, the term $(1-\tilde{\mu})\Pi^0 + \tilde{\mu}\Pi^1$ decreases. The same happens if we increase p_ν until it becomes active in Π^0. Hence the maximum of $(1-\tilde{\mu})\Pi^0 + \tilde{\mu}\Pi^1$ can be attained only if the inactive weights in Π remain inactive in both Π^0 and Π^1, and we may set $p_\nu = \tilde{\mu}$ for all these $\nu \notin A$.

Next, we claim that for a model $\nu \in A\setminus\{\mu\}$, which is active in Π, the maximum of $(1-\tilde{\mu})\Pi^0 + \tilde{\mu}\Pi^1$ can be only attained if ν remains active in both Π^0 and Π^1. To show this, we only need to argue that, regardless of the configuration of the other $p_{\nu'}$ $(\nu' \neq \nu)$,

$$\text{there is an assignment } p_\nu \in [0,1] \text{ such that both } \nu \in A^0 \text{ and } \nu \in A^1 \tag{22}$$

holds. If we then increase p_ν until (possibly) $\nu \notin A^1$, then we have that $(1-\tilde{\mu})\Pi^0 + \tilde{\mu}\Pi^1$ must decrease, since its derivative is smaller than zero.

We have that each p_ν in the interval $I^{01} := (1-\frac{1}{\tilde{w}_\nu}e^{-L^0(A^0)}, \frac{1}{\tilde{w}_\nu}e^{-L^1(A^1)})$ also satisfies (22). In order to show that I^{01} is non-empty, we first argue that $I^{01} \supset I := (1-$

$\frac{1}{\tilde{w}_\nu}e^{-L^0(A)}, \frac{1}{\tilde{w}_\nu}e^{-L^1(A)})$, which is then proven to be non-empty. Since we know that ν is active in Π and therefore $\tilde{w}_\nu < e^{-L(A)}$,

$$\frac{1}{e^{\sum_A \frac{\tilde{w}_\nu(1-p\nu)}{\tilde{w}^0(A)} \log \frac{1}{\tilde{w}_\nu(1-p\nu)}}} + \frac{1}{e^{\sum_A \frac{\tilde{w}_\nu p\nu}{\tilde{w}^1(A)} \log \frac{1}{\tilde{w}_\nu p\nu}}} = e^{-L^0(A)} + e^{-L^1(A)} \geq e^{-L(A)} \tag{23}$$

implies that I is not empty. We will verify (23) below.

$I \subset I^{01}$ holds by the following argument. Assume that for some $\nu' \in A$, $p_{\nu'}$ is so small that $\nu' \notin A^0$. Varying $p_{\nu'}$ in the range $[0, u]$ where $\nu' \notin A^0$, does not change the left constraint $1 - \frac{1}{\tilde{w}_\nu}e^{-L^0(A^0)}$, while the right constraint $\frac{1}{\tilde{w}_\nu}e^{-L^1(A^1)}$ is minimal at both boundaries $p_{\nu'} = 0$ and $p_{\nu'} = u$. This can be seen by considering the derivative $\frac{dL^1(A^1)}{dp_{\nu'}} = \frac{\tilde{w}_\nu}{\tilde{w}^1(A^1)}\big[-\log(\tilde{w}_{\nu'}p_{\nu'}) - L^1(A^1) - 1\big]$, which is $+\infty$ at $p_{\nu'} = 0$ and steadily decreases until $-\frac{\tilde{w}_\nu}{\tilde{w}^1(A^1)}(L^1(A^1)+1)$ at $p_{\nu'} = u$. Note that for both boundary points 0 and u, the value of $L^1(A^1)$ coincides. Thus we can set $p_{\nu'} = u$, making the interval I^{01} smaller. Letting $\tilde{A}^0 = A^0 \cup \{\nu'\}$ and $\tilde{A}^1 = A^1 \cup \{\nu'\}$, we then have $I^{01} = (1 - \frac{1}{\tilde{w}_\nu}e^{-L^0(\tilde{A}^0)}, \frac{1}{\tilde{w}_\nu}e^{-L^1(\tilde{A}^1)})$. A symmetric argument holds for the case that $\nu' \notin A^1$. In this way, we can subsequently treat all $\nu' \in A \setminus (A^0 \cap A^1)$, constantly decreasing I^{01}, until we arrive at I.

Now, in order to show (23), observe that $e^{\sum_A \frac{\tilde{w}_\nu(1-p\nu)}{\tilde{w}^0(A)} \log \frac{1}{(1-p\nu)}} \leq \frac{\tilde{w}(A)}{\tilde{w}^0(A)}$ and $e^{\sum_A \frac{\tilde{w}_\nu p\nu}{\tilde{w}^1(A)} \log \frac{1}{p\nu}} \leq \frac{\tilde{w}(A)}{\tilde{w}^1(A)}$ by Jensen's inequality, so (23) follows from

$$\frac{\tilde{w}^0(A)}{\tilde{w}(A)}e^{\sum_A \frac{\tilde{w}_\nu(1-p\nu)}{\tilde{w}^0(A)} \log \tilde{w}_\nu} + \frac{\tilde{w}^1(A)}{\tilde{w}(A)}e^{\sum_A \frac{\tilde{w}_\nu p\nu}{\tilde{w}^1(A)} \log \tilde{w}_\nu} \geq e^{\sum_A \frac{\tilde{w}_\nu}{\tilde{w}(A)} \log \tilde{w}_\nu}.$$

Applying Jensen's inequality again to the l.h.s. verifies this. Altogether we have shown so far that the maximum of $(1-\tilde{\mu})\Pi^0 + \tilde{\mu}\Pi^1$ can be only attained if $A = A^0 = A^1$.

Finally, we can turn to proving (21), by showing that the maximum of $(1-\tilde{\mu})\Pi^0 + \tilde{\mu}\Pi^1$ is attained if $p_\nu = \tilde{\mu}$ for all $\nu \in \mathcal{C}$. Since we know already that we may set $p_\nu = \tilde{\mu}$ for $\nu \notin A$ in order to attain the maximum, we can just ignore these models and assume without loss of generality that $A = \mathcal{C}$. Then the derivatives of Π^0 and Π^1 are

$$\begin{aligned} \frac{d\Pi^0}{dp_\nu} &= -\frac{\tilde{w}_\nu}{w^0(A)}\big(-\log[\tilde{w}_\nu(1-p_\nu)] - L^0(A^0)\big) \text{ and} \\ \frac{d\Pi^1}{dp_\nu} &= \frac{\tilde{w}_\nu}{w^1(A)}\big(-\log[\tilde{w}_\nu p_\nu] - L^1(A^1)\big), \end{aligned}$$

respectively. A possible maximum has $(1-\tilde{\mu})\frac{d\Pi^0}{dp_\nu} + \mu\frac{d\Pi^1}{dp_\nu} = 0$ for all $\nu \neq \mu$, which occurs in case that $p_\nu = \tilde{\mu}$ for all $\nu \in \mathcal{C}$. This is in fact a global maximum if we can show the Hessian is globally negative semi-definite. It is sufficient to show that both Hessians of Π^0 and Π^1 are negative semi-definite: We identify the model class with an index set $\mathcal{C} = A \cong \{0, 1, \ldots, N\}$ and assign the true distribution to the index 0. Then, abbreviating $D_i = \log(\tilde{w}_i p_i) - L^1(A)$ and using the characteristic function $\mathbb{1}_{i=j}$ which is one if $i = j$ and zero otherwise, the Hessian of Π^1 is computed as

$$\left[\frac{d^2\Pi^1}{dp_i dp_j}\right]_{i,j=1}^N = -\frac{1}{\tilde{w}^1(A)^2}\left[\tilde{w}_i\tilde{w}_j\big(\mathbb{1}_{i=j}\frac{\tilde{w}^1(A)}{\tilde{w}_i} + D_i + D_j - 1\big)\right]_{i,j=1}^N.$$

This Hessian is negative semi-definite by Lemma 15 below, and so is the Hessian of Π^0. This concludes the proof. □

Lemma 15 *Let $N \geq 1$ and $w_i > 0$ for $0 \leq i \leq N$ (the w_i need not sum up to one). Let $W = \sum_{i=0}^N w_i$ and assume that $-\log w_j \geq L := -\sum_{i=0}^N \frac{w_i}{W} \log w_i$ holds for all $1 \leq j \leq N$. Then, for all vectors $u \in \mathbb{R}^N$, we have that*

$$\sum_{i,j=1}^N u_i u_j \left[\frac{\mathbb{1}_{i=j} W}{w_i} - \log w_i - L - \log w_j - L - 1 \right] \geq 0. \tag{24}$$

Proof. We proceed by induction over N. For $N = 1$, the assertion is immediate. Now, for N, observe that the derivative of the l.h.s. of (24) w.r.t. w_0,

$$\sum_{i=1}^N \frac{u_i^2}{w_i} + \frac{2\big(\sum_{i=1}^N u_i\big)^2}{W} \Big[1 + L + \log w_0 \Big],$$

is positive, since $-\log w_0 - L < 0$. Thus we may decrease the l.h.s. of (24) by decreasing w_0, until eventually $-\log w_k = L$ holds for one $k \in \{1 \dots N\}$. Set $D_i = -\log w_i - L$ and $\tilde{W} = W - w_k$. Then

$$\begin{aligned} \sum_{i,j=1}^N u_i u_j \left[\frac{\mathbb{1}_{i=j} W}{w_i} + D_i + D_j - 1 \right] &= \sum_{i,j \in \{1 \dots N\} \setminus \{k\}} u_i u_j \left[\frac{\mathbb{1}_{i=j} \tilde{W}}{w_i} + D_i + D_j - 1 \right] \\ &+ \sum_{i \in \{1 \dots N\} \setminus \{k\}} \left[\frac{w_i}{w_k} u_k^2 - 2(1 - D_i) u_i u_k + \frac{w_k}{w_i} u_i^2 \right]. \end{aligned} \tag{25}$$

Since for all $u, v \in \mathbb{R}$ and $c \leq 1$ we have $u^2 - 2cuv + v^2 \geq 0$, the term (25) is nonnegative. Thus the assertion follows from the induction hypothesis. □

Now we can prove the main result of this paper.

Theorem 16 (Predictive performance of stochastic model selection) *For countable model class $\mathcal{C}$ containing the true distribution μ, define the potential as*

$$\mathcal{P}(h_{<t}) = \big[\mathcal{K}(h_{<t}) + \log(1 + \mathcal{H}(h_{<t})) \big] (1 + \Pi(h_{<t})).$$

Then, for any history $h_{<t}$ and current input z_t,

$$\begin{aligned} &\mathcal{P}(h_{<t}) - \mathbf{E}_{x_t \sim \mu(\cdot|z_t)} \mathcal{P}(h_{1:t}) \geq \mathcal{H}(h_{<t}) - \mathbf{E}_{x_t \sim \xi(\cdot|z_t, g_{<t})} \mathcal{H}(h_{1:t}), \text{ and thus} \\ &\sum_{t=1}^\infty \mathbf{E} \big\| \Xi - \xi \big\|_2^2 \leq \mathcal{P}^{\text{init}} = (1 + \Pi^{\text{init}}) \big[\log(1 + \mathcal{H}^{\text{init}}) - \log(w_\mu^{\text{init}}) \big]. \end{aligned} \tag{26}$$

Proof. Using Theorem 14, this follows directly from Lemma 9. □

Theorem 16 implies performance guarantees for arbitrary bounded loss functions.

Corollary 17 *For each input z, let $\ell(\cdot,\cdot|z) : (\hat{x},x) \mapsto \ell(\hat{x},x|z) \in [0,1]$ be a loss function known to the learner, depending on the true outcome x and the prediction $\hat{x}$ (ℓ may also depend on the time, but we don't complicate notation by making this explicit). Let $\ell^{\mu}_{<\infty}$ be the cumulative loss of a predictor knowing the true distribution μ, where the predictions are made in a Bayes optimal way (i.e. choosing the prediction $\arg\min_{\hat{x}} \mathbf{E}_{x\sim\mu}\ell(\hat{x},x|z_t)$ for current input z_t), and $\ell^{\Xi}_{<\infty}$ be the corresponding quantity for the stochastic model selection learner. Then the loss of the learner is bounded by*

$$\mathbf{E}\ell^{\Xi}_{<\infty} \leq \mathbf{E}\ell^{\mu}_{<\infty} + 2C + 2\sqrt{2C\mathbf{E}\ell^{\mu}_{<\infty}},$$

where $C = \left(\sqrt{(1+\Pi^{\text{init}})\left[\log(1+\mathcal{H}^{\text{init}}) + \mathcal{K}^{\text{init}}\right]} + \sqrt{\mathcal{K}^{\text{init}}}\right)^2$.

Proof sketch. First, we have to prove Theorem 16 for the Hellinger distance instead of the quadratic distance, arriving at $\sum_{t=1}^{\infty}\mathbf{E}\left\|\sqrt{\Xi}-\sqrt{\mu}\right\|_2^2 \leq C$. This is straightforward, since Lemma 4 and Lemma 6 also hold for the Hellinger distance. Then, the result follows from [PH05, Lemma 24–26] with one additional application of Jensen's inequality. □

Similar and even slightly stronger loss bounds hold for different stochastic model selection algorithms derived from prediction with expert advice, e.g. the Hedge algorithm [FS97] for finite model class or FPL [HP05] for countable class. The expert proof techniques do not even require $\mu \in \mathcal{C}$, but work without any assumption on the data generating process. The experts posterior, according to which is sampled, is not Bayesian, but specifically designed in order to minimize loss. However, these algorithms do not give estimates for the true probabilities μ, and consequently no consistency in this sense can be proved. We expect that Bayesian algorithms are superior to expert algorithms in cases where probability estimates can be obtained and are beneficial.

5 The magnitude of the entropy potential

In this section, we will partially answer the question how large the newly defined quantity, the entropy potential, can grow. We start with a general bound.

Proposition 18 *The μ-entropy potential is always bounded by $\Pi \leq \mathcal{H}/w_\mu$. There are cases where this bound is sharp up to a factor.*

Proof. With A denoting the active set, we have that $\mathcal{H} \geq -\sum_{\nu\in A} w_\nu \log w_\nu = w(A)L(A) \geq w_\mu L(A) \geq w_\mu\Pi$. In order to see that this bound is sharp in general, consider the model class $\mathcal{C} = \{\mu,\nu^*,\nu_1\ldots\nu_N\}$, where $w_{\nu^*} \geq \frac{1}{2}$, $w_\mu = \frac{1}{2}(1-w_{\nu^*})$, and $w_{\nu_1} = \ldots = w_{\nu_N} = \frac{1}{2N}(1-w_{\nu^*})$. Then $\Pi > \log 2 + \frac{1}{2}\log N$, but $\mathcal{H} = H(w_{\nu^*},1-w_{\nu^*}) + 2w_\mu(\log 2 + \frac{1}{2}\log N)$. Thus, $\Pi > \frac{1}{2}w_\mu^{-1}(\mathcal{H} - H(w_{\nu^*},1-w_{\nu^*}))$ holds in this case. □

Proposition 18 gives a worst-case bound which is of course not satisfactory: Using it in Theorem 16, the resulting bound becomes no better than that of Theorem 7. Fortunately, in most cases, the entropy potential behaves well and is much smaller. For finite model classes, it is clearly at most the logarithm of the size of the class. We present two different infinite model classes in the following, one where the weights decay rapidly, and one where they decay slowly.

Example 19 Consider a model class with rapidly decaying weights $(2^{-i})_{i\geq 1}$, and suppose that the true model has index k. Then it has complexity $k \log 2$. Moreover, since $\sum_{i=k}^{\infty} \frac{i \log 2}{2^{i-k+1}} = (k+1)\log 2$, in the computation of Π, all models $i \geq k$ are active, and consequently $\Pi = \log(k+3) = O(\mathcal{K})$. In contrast, the entropy of the model class is $2 \log 2$.

Example 20 Consider a model class with slowly decaying weights $(\frac{6}{i^2\pi^2})_{i\geq 1}$, and let the true model have index k, i.e. complexity $\log \frac{k^2\pi^2}{6}$. In order to estimate Π this time, we have to find a sufficiently large index j such that $\log \frac{j^2\pi^2}{6} \geq \big[\frac{1}{k^2} \log \frac{k^2\pi^2}{6} + \sum_j^{\infty} \frac{1}{i^2} \log \frac{i^2\pi^2}{6}\big] / [\frac{1}{k^2} + \sum_j^{\infty} \frac{1}{i^2}]$. Substituting the sums by appropriate integrals, we make the r.h.s. larger and look for a j such that

$$\log \frac{j^2\pi^2}{6} \left[\tfrac{1}{k^2} + \tfrac{1}{j}\right] \geq \tfrac{1}{k^2} \log \frac{k^2\pi^2}{6} + \frac{\log \frac{(j-1)^2\pi^2}{6} + 2}{j-1}.$$

This is satisfied for $j - 1 = k^2$, as an elementary computation verifies. Consequently, the active set consists of at most k^2 models, and $\Pi \leq \log\big(k^2 + O(k^2 \log k^4)\big) = O\big(\log(k)\big) = O(\mathcal{K})$. The entropy of the model class is $O(1)$.

In both of the above examples, we have that $\Pi = O(\mathcal{K}) = O(-\log w_\mu)$. We believe that this is the *typical* behavior of the entropy potential, in contrast to Proposition 18. It remains open to precisely characterize those well-behaved cases:

Problem 21 *Characterize those weight distributions satisfying* $\Pi = O(\mathcal{K})$.

Finally, our bounds are infinite with the usual definition of a *universal model class*. But with a slight modification of the prior, they become finite. Hence we can obtain a universal induction result for stochastic model selection:

Example 22 Consider a model class $\mathcal{C}$ corresponding to the set of programs on a universal Turing machine. For $\nu \in \mathcal{C}$, let $w_\nu \sim 2^{-K(\nu)}/K(\nu)^2$, where K denotes the prefix Kolmogorov complexity – it is shown e.g. in [LV97] how to obtain such a construction. Then $\mathcal{H} = O(1)$, and Theorem 16 implies consistency of universal stochastic model selection with this prior and normalization. If we would have chosen the usual "canonical" weights $w_\nu \sim 2^{-K(\nu)}$, then $\mathcal{H} \cong \sum K(\nu) 2^{-K(\nu)} = \infty$, since K is the smallest possible code length to satisfy the Kraft inequality, and any smaller code length must necessarily result in an infinite sum. Hence the bound for universal stochastic model selection is infinite with the usual prior.

6 Discussion and conclusions

We have shown that the cumulative quadratic error of Bayesian stochastic model selection in countable model classes is finitely bounded if the entropy is so. This corresponds to results for Bayes mixture and MDL learners (which however have no entropy in their bounds). Unlike the MDL bound [PH05], our bound obtained here is usually not exponential in the complexity of the true distribution (but can be so in bad cases). Our results imply strong consistency (almost sure convergence) of the stochastic model selection learner, and moreover loss bounds for arbitrary bounded loss functions. According to Examples 1, 8, and 11, our cumulative error bound (26), which is of order $\mathcal{K} \cdot \Pi$, is essentially sharp in the sense that both ingredients $\mathcal{K}$ and Π are necessary. We don't know if the bound needs to be the product $\mathcal{K} \cdot \Pi$, or if the smaller sum $\mathcal{K} + \Pi$ is also possible.

There is hope that the entropy potential introduced in this work has other applications in learning and information theory. It remains open to study this quantity more thoroughly and characterize the weight distributions where it behaves well (Problem 21). Moreover, this paper show new ways to evaluate the state of learning in a discrete model class. The general potential $\mathcal{P}$ from Theorem 16 always strictly decreases in expectation unless all models predict the same (so it does not share the undesirable property of $\mathcal{K}$ discussed in Example 5). A couple of interesting questions remain open, such as defining a good potential function for active learning.

References

[CB90] B. S. Clarke and A. R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inform. Theory*, 36:453–471, 1990.

[CBL03] N. Cesa-Bianchi and G. Lugosi. Potential-based algorithms in on-line prediction and game theory. *Machine Learning*, 51(3):239–261, 2003.

[FS97] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[HP05] M. Hutter and J. Poland. Adaptive online prediction by following the perturbed leader. *Journal of Machine Learning Research*, 6:639–660, 2005.

[Hut04] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2004.

[LV97] M. Li and P. M. B. Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer, 2nd edition, 1997.

[McA03] D. McAllester. PAC-bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.

[PH05] J. Poland and M. Hutter. Asymptotics of discrete MDL for online prediction. *IEEE Transactions on Information Theory*, 51(11):3780–3795, 2005.

[Ris96] J. J. Rissanen. Fisher Information and Stochastic Complexity. *IEEE Trans. Inform. Theory*, 42(1):40–47, January 1996.

[Sol78] R. J. Solomonoff. Complexity-based induction systems: comparisons and convergence theorems. *IEEE Trans. Inform. Theory*, 24:422–432, 1978.