𝕋ℂ𝕊-𝕋ℝ-𝔸-**06-23**

𝕋ℂ𝕊 Technical Report

Consistency Theorems for Discrete Bayesian Learning

by

J. Poland

Division of Computer Science

Report Series A

November 30, 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-06-23 November 30, 2006

Consistency Theorems for Discrete Bayesian Learning

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

Bayes' rule specifies how to obtain a posterior from a class of hypotheses endowed with a prior and the observed data. There are three fundamental ways to use this posterior for predicting the future: marginalization (integration over the hypotheses w.r.t. the posterior), MAP (taking the a posteriori most probable hypothesis), and stochastic model selection (selecting a hypothesis at random according to the posterior distribution). If the hypothesis class is countable and contains the data generating distribution (this is termed the "realizable case"), strong consistency theorems are known for the former two methods, asserting almost sure convergence of the predictions to the truth as well as loss bounds. We prove corresponding results for stochastic model selection, for both discrete and continuous observation spaces. As a main technical tool, we will use the concept of a potential: this quantity, which is always positive, measures the total possible amount of future prediction errors. Precisely, in each time step, the expected potential decrease upper bounds the expected error. We introduce the *entropy potential* of a hypothesis class as its worst-case entropy with regard to the true distribution. Our results are proven within a general stochastic online prediction framework that comprises both online classification and prediction of non-i.i.d. sequences.

1 Introduction

"When you have eliminated the impossible, whatever remains must be the truth." This famous quote describes the induction principle of Sherlock Holmes, whose observations and conclusions are always correct. Real world observations usually lack this desirable

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

property, instead they are *noisy.* Thus, Bayes' rule, *eliminating the improbable*, has emerged as a successful induction principle in practice. The aim of this paper is to collect and prove statements of the form: "When you have eliminated the improbable, whatever remains is almost sure to behave like the truth." We will give different but tightly connected forms of this assertion: Asymptotic almost sure consistency results and bounds on the error of a predictor based on Bayes' rule.

1.1 Structure and contributions of this work

The main technical contribution of this paper, presented in Section 3, are several proofs of consistency theorems for Bayesian stochastic model selection. This completes a series of recent performance guarantees obtained for all three fundamental ways of Bayesian learning. It therefore motivates a comparative presentation of all these results, discussing the basics of Bayesian learning, the fundamental variants of Bayesian induction, its scope of applicability, and the state of the art of Bayesian learning theorems. This is subject of the next section.

2 Discrete Bayesian Learning

Bayes' famous rule,

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}, \tag{1}$$

says how the probability of a hypothesis H is updated after observing some data D. Still, different specific induction setups can use Bayes' rule. First, there are different possibilities to define the input space, the observation space, and the hypothesis space. Second, a hypothesis class endowed with a probability distribution can be used for induction in principally three different ways.

The reader should keep in mind that Bayes' rule is no theorem in general. Under the assumption that hypotheses and data are *both* sampled from a joint probability distribution that coincides with the prior $P(H)$, (1) would be a theorem. However, Bayes' rule is commonly not applied under such an assumption, in particular the distribution $P(H)$ on the hypotheses is usually merely a *belief distribution*, there is no probabilistic sampling mechanism generating hypotheses assumed. Hence, Bayes' rule is motivated intuitively in the first place. Still, many optimality results and performance guarantees have been shown for Bayesian induction (e.g. in [BD62, CB90, BRY98]), including the results of the present work.

2.1 What to learn? Hypotheses, history, inputs, observation spaces

Let $\mathcal{X}$ be the observation space. We work in an *online prediction setup in discrete time*, that is, in each time step $t = 1, 2, \ldots$, an observation $x_t \in \mathcal{X}$ is revealed to the learner. The task of the learner will be to predict x_t *before* he observes it. One question of fundamental technical impact concerns the structure of the observation space $\mathcal{X}$. We restrict our attention to the two most important cases of (a) $\mathcal{X}$ being discrete (finite or countable) and (b) continuous $\mathcal{X} \subset \mathbb{R}^d$ for suitable dimension $d \in \mathbb{N}$. As any discrete space can be mapped to a subset of $\mathbb{R}^d$, it is technically sufficient to restrict to $\mathcal{X} \subset \mathbb{R}^d$, which we will do in the following (except for a few places where we explicitly deal with finite observation spaces).

A *hypothesis* ν specifies a probability distribution on the observation space $\mathcal{X}$. In the simplest case, it does not depend on any input, these hypotheses represent the assumption that the observed data is independently identically distributed (i.i.d.). In all other cases, there is some *input space* $\mathcal{Z}$, and a hypothesis maps inputs to distributions on $\mathcal{X}$. In fact, technically, the inputs play no role at all, as we will see in the following. We therefore may assume the existence of an arbitrary input space $\mathcal{Z}$ without any structure (which may consist of just one point, meaning that there are no inputs at all), and inputs are generated by an arbitrary process. This covers (even more than) two of the most important learning setups: *Classification*, where the data is conditionally i.i.d. given the inputs, and *prediction of non-i.i.d. sequences*, where in each time step t, we may define the input $z_t = (x_1, \ldots, x_{t-1})$ to be the observation history seen so far. Generally, we will denote the history of inputs *and* observations by

$$h_{1:t-1} = h_{<t} = (z_1, x_1, z_2, x_2, \ldots, z_{t-1}, x_{t-1})$$

(observe that two pieces of notation have been introduced here).

Now, a hypothesis is formally defined as a function

$$\nu : \mathcal{Z} \to \mathcal{M}^1_{\mathrm{D+C}}(\mathcal{X}).$$

Here, $\mathcal{M}^1_{\mathrm{D+C}}(\mathcal{X})$ denotes the probability distributions on $\mathcal{X} \subset \mathbb{R}^d$, that are mixtures of discrete distributions (with nonzero mass concentrated on single points) and distributions with continuous density functions. We make this restriction mainly because we wish to be able to define all subsequent quantities, in particular Bayesian posteriors, effortlessly[1] and uniquely (except perhaps on a set of measure zero).[2] In particular, we have

$$\int d\nu(\cdot|z) = 1 \text{ for all } z \in \mathcal{Z}.$$

Note that we consistently use this integral notation, also for discrete observation space (in which case the integral is in reality a sum).

[1] For instance, for a measure defined by a "devil's staircase", one has to spend additional effort in order to define everything properly, which is not the aim of the present work. However, this and other cases can be treated with the methods described here.

[2] The continuity assumption will be needed for the main proof in Section 3. It can be immediately lifted and replaced by "uniform piecewise continuity", which means that there is a single partition of $\mathcal{X}$ such that the continuous parts of all distributions $\nu \in \mathcal{C}$ and for all $z \in \mathcal{Z}$ are continuous on each of the elements of the partition. Maybe it can be even further lifted.

A Bayesian learner is always based on a *hypothesis class* $\mathcal{C} = \{\nu_1, \nu_2, \ldots\}$. In this work with the title "discrete Bayesian learning", we restrict to discrete, i.e. finite or countable, hypothesis classes (and in the notation we assume a countable hypothesis class from now on, without loss of generality). Before the learning process starts, each hypothesis $\nu \in \mathcal{C}$ is endowed with a prior weight $w_\nu \in (0, 1)$, such that $\sum_{\nu \in \mathcal{C}} w_\nu = 1$.

Hypothesis classes considered in statistics are usually *continuously parameterized*. One motivation to study discrete classes is that they are technically simpler, so they can serve as a basis for the more advanced continuous case. In the continuous case, some Bayesian predictors such as MAP (see below) are not consistent at all, while others such as MML (minimum message length) [WB68, WD99] and MDL (minimum description length) [Ris96] require appropriate discretization. Also, countable hypothesis classes always admit stronger performance guarantees than possible for their continuously parameterized counterparts. In particular, we will be able to show almost sure consistency, whereas only convergence in probability holds in the continuous case (e.g. in [BC91]).

Another particular motivation to consider discrete hypothesis classes arises in Algorithmic Information Theory. General continuous hypothesis classes are computationally not tractable. The largest hypothesis class which can be manipulated in the limit by a computer, is the class of all computable hypotheses on some fixed universal Turing machine, precisely prefix machine [LV97]. Thus each hypothesis corresponds to a program, and there are countably many programs. Each hypothesis has a natural description length, namely the length of the corresponding program. If we agree that programs are binary strings, then a natural prior is defined by two to the negative description length.

If we are dealing with such a *universal* hypothesis class as defined in Algorithmic Information Theory, we need to be careful about the phenomenon of *probability leaks*: A hypothesis, that is a program on our universal Turing machine, may not produce output for certain inputs. Because of our inability of deciding the halting problem, we cannot generally detect this case. As a consequence, there is no limit-computable way of defining hypotheses that are proper probability distributions, they are rather *semimeasures*. In this paper, we won't address this issue further, instead we point to the references: Consistency theorems for the semimeasure case are known for marginalization [Sol78, Hut04] and for MAP predictions [PH05], but not for stochastic model selection. All of the probability distributions considered in this paper will be proper measures.

We rewrite Bayes' rule (1) using new notation: For a hypothesis $\nu \in \mathcal{C}$, current prior weights $w_{\nu'}(h_{<t})$ of all hypotheses $\nu' \in \mathcal{C}$ depending on the history $h_{<t}$, input z_t, and observation x_t, we set the posterior weight of ν to

$$w_\nu(h_{1:t}) = \frac{\nu(x_t|z_t) \cdot w_\nu(h_{<t})}{\sum_{\nu' \in \mathcal{C}} \nu'(x_t|z_t) \cdot w_{\nu'}(h_{<t})}, \tag{2}$$

Note that we actually need to distinguish three variants of Bayes' rule (not to be confused with the three variants of Bayesian prediction discussed below): In the case of discrete observation space, the quantities $\nu'(x|z)$ (and therefore also the sum in the denominator) are probabilities, while for continuous observation space, they are densities. Finally, if at least one hypothesis $\nu \in \mathcal{C}$ is a mixture of a discrete and a continuous distribution, then *all* $\nu'(x|z)$ must be treated as mixtures in the following way: If for an observation $x \in \mathcal{X}$, there is a hypothesis assigning non-zero mass to x, then the $\nu'(x|z)$ are treated as probabilities (and all hypotheses assigning merely a non-zero density to that particular x will get posterior weight 0). Otherwise, the $\nu'(x|z)$ are treated as densities.

2.2 How to learn? Three fundamental variants of Bayesian prediction

Given a set of hypotheses $\mathcal{C}$ and some observed data $h_{1:t} = (z_1, x_1, \dots, z_t, x_t)$, a legitimate question is asking which of the hypotheses in $\mathcal{C}$ has actually generated the data. It is clear that this question might not be well-defined if the process generating the data, which we will call μ in the sequel, is *not* member of $\mathcal{C}$. Actually, one can immediately construct examples where any Bayesian learner produces very undesirable results in this non-realizable learning setup (see [GL04] for sophisticated examples). In this work, we will restrict to the *realizable* case, where the true distribution generating the observations is contained in the class, $\mu \in \mathcal{C}$. (But recall that this only refers to the distribution of the observation *given* the inputs, we do not need any assumption on the generation of the inputs z_t). Of course, the learner does not know in advance which element of $\mathcal{C}$ is the true distribution μ.

However, hypothesis identification has technical difficulties. For instance, consider the case where two hypothesis are in $\mathcal{C}$ that make (almost) identical predictions, one of them being the true one. Then it is (almost) impossible to identify the right one, but if we just want to make predictions, we need not care: Choosing any of the two will yield (almost) perfect predictions.

So from now on, we restrict our focus to prediction. That is, for given history $h_{<t}$ and current input $z_t \in \mathcal{Z}$, we are interested in a *predictive distribution*[3] on the observation space $\mathcal{X}$ that comes as closely to the truth as possible. Our hypothesis class endowed with the Bayesian posterior $\big(w_{\nu'}(h_{<t})\big)_{\nu' \in \mathcal{C}}$ offers us *three fundamental* ways to obtain such a prediction:

1. **Marginalization**. If we apply Bayes' rule (1) to the modified setting where the next observation x_t takes the place of the hypothesis H, then, as an easy computation shows, we get a predictive distribution $\xi(x_t|z_t, h_{<t})$ by integrating the predictions of *all* hypotheses w.r.t. the current posterior:

$$\xi(x|z_t, h_{<t}) = \sum_{\nu' \in \mathcal{C}} w_{\nu'}(h_{<t}) \nu'(x|z_t). \tag{3}$$

[3] In many prediction tasks, a single value is required as prediction, rather than a distribution. Such a single prediction can be derived from a predictive distribution, e.g. by minimizing a risk function, compare Corollary 5 below.

2. **Maximum a posteriori (MAP).** If we are interested in a *single hypothesis' prediction*, then we may choose the *hypothesis with maximal a-posteriori belief value*, abbreviated as MAP hypothesis:

$$\nu^*_{h_{<t}} = \arg\max_{\nu\in\mathcal{C}}\{w_\nu(h_{<t})\} \text{ and} \tag{4}$$

$$m(x_t|z_t,h_{<t}) = \nu^*_{h_{<t}}(x_t|z_t), \tag{5}$$

where the latter $m(x_t|z_t,h_{<t})$ is the MAP prediction.

3. **Stochastic model selection.** The third possibility is to randomize and sample a hypothesis according to the probability distribution defined by the current posterior. This *stochastic model selection* can be formally written as

$$\Xi(x_t|z_t,h_{<t}) = \tilde{N}(x_t|z_t) \text{ where } \tilde{N}\in\mathcal{C} \tag{6}$$
$$\text{and } \mathbf{P}(\tilde{N}=\nu') = w_{\nu'}(h_{<t}) \text{ for all } \nu'\in\mathcal{C}.$$

Note that for given history $h_{<t}$, the first two methods are deterministic, resulting in a fixed predictive distribution. Stochastic model selection uses additional randomness.

There are other possibilities than the stated three to use a Bayesian hypothesis class for prediction. MAP is tightly related to MML and MDL, but the terms MML and MDL are (also) used for (slightly, in the case of discrete hypothesis class) different concepts [CD05, Ris96]. Also, there is a "dynamic" variant of MAP defined in [PH05], where a MAP hypothesis is chosen for each possible outcome x_t and used for prediction. Anyway: many, if not most, Bayesian prediction methods can be roughly grouped into the three fundamental "integrate over all hypothesis", "take the hypothesis with the best current score", and "select one hypothesis at random according to the current belief distribution". And we hold (but that is a matter of taste) that the above representants are the simplest and most natural of the prediction methods to consider.

2.3 Performance guarantees for Bayesian learners

We are now ready to state the performance guarantees for the three Bayesian learners defined in (3), (5), and (6). We start with the technically easiest case of marginalization (3). Actually, this result has been originally discovered by Solomonoff [Sol78] within the context of Algorithmic Information Theory.

Recall that $\mu\in\mathcal{C}$ is the true distribution generating the data, and ξ is the marginalization predictor. The quadratic *Hellinger distance* between the ξ-predictions and μ-predictions at time t is given by

$$\Delta^2_t(\mu,\xi) := \int d\big(\sqrt{\mu(\cdot|z_t)} - \sqrt{\xi(\cdot|z_t,h_{<t})}\big)^2. \tag{7}$$

It clearly depends on the history $h_{<t}$ and the current input z_t. Our main technical results are all stated as cumulative (i.e., over $t=1,\ldots,\infty$) bounds on the Hellinger distance (that is, *errors*) of the predictive probabilities to the truth.

Theorem 1 *If $\mu \in \mathcal{C}$, then for* any *sequence of inputs $z_1, z_2, \ldots$,*

$$\sum_{t=1}^{\infty} \mathbf{E}_\mu \Delta_t^2(\mu, \xi) \leq \log w_\mu^{-1} \tag{8}$$

holds, where log *denotes the natural logarithm and w_μ is the prior weight of the true distribution. $\mathbf{E}_\mu$ refers to the fact that the expectation is taken w.r.t. the true distribution μ, i.e., all observations are generated w.r.t. μ conditional to the inputs, and this expectation is computed.*

Remark 2 *The following convention is used throughout the paper: When we write w_μ, it always refers to the prior weight $w_\mu(\varnothing)$, with empty history. Posterior weights are being denoted with explicit history, $w_\mu(h_{<t})$. This convention will extend to other quantities, such as the entropy $\mathcal{H}$ or the entropy Π below. However, in the proofs below, the history $h_{<t}$ at a given time t is often dropped for notational convenience. In these cases, a notation like w_μ refers to the current posterior weight.*

It should not be surprising that the quantity w_μ appears on the r.h.s. and therefore has an impact on how large the error on the l.h.s. can grow. After all, if the Bayesian learner assigns a high prior weight to the true distribution, the error should be small. The remarkable fact is the *logarithmic* dependence in w_μ. As by Kraft's inequality, the logarithm of a weight can be interpreted as its description length, (8) is a very strong result asserting that the cumulative error never exceeds the description length of the true distribution. In a sense: When finding the truth single-handedly, our error is at most the number of bits a teacher needs to tell us the truth. We will provide a proof of Theorem 1 at the beginning of Section 3, as an introduction for the subsequent proof techniques.

Results for the MAP predictor (5) similar to Theorem 1 have been shown in [PH05].

Theorem 3 *Assume $\mu \in \mathcal{C}$. Suppose that, for any history with nonzero probability density, the hypotheses always admit the specification of a (not necessarily unique) MAP hypothesis ν^*. This is satisfied for instance if all hypotheses correspond to continuous probability densities that are uniformly bounded. Then*

$$\sum_{t=1}^{\infty} \mathbf{E}_\mu \Delta_t^2(\mu, m) \leq 21 w_\mu^{-1}. \tag{9}$$

The proof uses telescoping and dominance. The most remarkable (and worrying) fact here is the bound $O(w_\mu^{-1})$ on the r.h.s. While the logarithm in (8) is sufficiently small to be of practical significance, the exponentially larger quantity $O(w_\mu^{-1})$ is generally huge. One can construct examples where this bound is sharp [PH06]. Fortunately, this does not necessarily imply that the MAP predictions are bad, the actual error is smaller in many important cases. Still, there are situations where

MAP predictions tend to be "unbalanced" and therefore unfavorable compared to marginalization. Stochastic model selection often gives better results in such cases.

The corresponding theorem for stochastic model selection (6), which is the main technical result of this paper, reads as follows.

Theorem 4 *Assume $\mu \in \mathcal{C}$. Then, for any sequence of inputs $z_1, z_2, \ldots$,*

$$\sum_{t=1}^{\infty} \mathbf{E}_\mu \mathbf{E}_\Xi \Delta_t^2(\mu, \Xi) = O\big(w_\mu^{-1} + \Pi(\log \mathcal{H} + \log w_\mu^{-1})\big) = O(\Pi \log w_\mu^{-1}) \tag{10}$$

holds. The quantities $\mathcal{H}$ and Π, the Shannon entropy and the μ-entropy potential of the hypothesis class, are defined below. $\mathbf{E}_\Xi$ serves as a reminder that the Ξ-predictor is randomized.

The quantity $\mathcal{H}$ in the theorem is the Shannon entropy of the hypothesis class w.r.t. the current posterior distribution,

$$\mathcal{H}(h_{<t}) = \mathcal{H}\big([w_\nu(h_{<t})]_{\nu \in \mathcal{C}}\big) = -\sum_{\nu \in \mathcal{C}} w_\nu(h_{<t}) \log w_\nu(h_{<t}).$$

According to the convention in Remark 2, if we write just $\mathcal{H}$ as in the theorem, this corresponds to the prior (or, below in the proofs, to the current posterior). Moreover, we define the current *entropy potential of the hypothesis class relative to the true distribution* μ as

$$\Pi\big((w_\nu)_{\nu \in \mathcal{C}}\big) = \sup\big\{\mathcal{H}\big((\tfrac{\tilde{w}_\nu}{\sum_{\nu'} \tilde{w}_{\nu'}})_{\nu \in \mathcal{C}}\big) : \tilde{w}_\mu = w_\mu \wedge \tilde{w}_\nu \leq w_\nu \ \forall \nu \in \mathcal{C} \setminus \{\mu\}\big\} \tag{11}$$

and $\Pi(h_{<t}) = \Pi\big([w_\nu(h_{<t})]_{\nu \in \mathcal{C}}\big)$, see also Definition 15 below. This can be paraphrased as "worst-case entropy of the class under all possible Bayesian updates where the true distribution always has evidence value 1". We use the same convention as before: writing just Π corresponds to the prior, or, in the proofs below, to the current posterior (Remark 2).

As we will see in Section 3.3, the entropy potential Π can grow as large as $\Omega(\mathcal{H} w_\mu^{-1})$ in general. However, if the prior $(w_\nu)_{\nu \in \mathcal{C}}$ has sufficiently *light tails*, Π is of order at most $\log w_\mu^{-1}$.

2.4 Implications: almost sure consistency and loss bounds

One important consequence of any finite bound on the expected cumulative Hellinger error is *almost sure consistency* of the predictor in the Hellinger sense. That is, the Hellinger distance of the predictive to the true distribution tends to zero almost surely. This is easily verified, as

$$\mathbf{P}\Big(\exists t \geq T : \Delta_t^2 \geq \varepsilon\Big) = \mathbf{P}\Big(\bigcup_{t \geq T} \{\Delta_t^2 \geq \varepsilon\}\Big) \leq \sum_{t \geq T} \mathbf{P}\big(\Delta_t^2 \geq \varepsilon\big) \leq \frac{1}{\varepsilon} \sum_{t=T}^{\infty} \mathbf{E}\Delta_t^2 \overset{T \to \infty}{\longrightarrow} 0 \tag{12}$$

holds. In case of a finite or countable observation space $\mathcal{X}$, this implies in particular convergence of all predictive probabilities $\xi(x_t|z_t, h_{<t})$ to the true probabilities $\mu(x_t|z_t)$. In case of a continuous observation space, the predicted probability masses of any measurable subset of $\mathcal{X}$ converges to the true mass. However, we cannot conclude the convergence of moments, e.g. the expectation, without making further assumptions.

Other implications of Theorems 1–4 are *loss bounds* of a Bayes-optimal decision maker based on the predictive distribution, w.r.t. arbitrary loss functions. The proof of the following corollary proceeds as that of [PH05, Theorem 27].

Corollary 5 *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 $\ell^\xi_{<\infty}$, $\ell^m_{<\infty}$, $\ell^\Xi_{<\infty}$ be the corresponding quantities for the respective Bayesian learners. Then the loss of the learners are bounded by*

$$\mathbf{E}\ell^\xi_{<\infty} \leq \mathbf{E}\ell^\mu_{<\infty} + O\big(\log w_\mu^{-1}\big) + O\Big(\sqrt{\log w_\mu^{-1}\mathbf{E}\ell^\mu_{<\infty}}\Big), \tag{13}$$

$$\mathbf{E}\ell^m_{<\infty} \leq \mathbf{E}\ell^\mu_{<\infty} + O\big(w_\mu^{-1}\big) + O\Big(\sqrt{w_\mu^{-1}\mathbf{E}\ell^\mu_{<\infty}}\Big), \text{ and} \tag{14}$$

$$\mathbf{E}\ell^\Xi_{<\infty} \leq \mathbf{E}\ell^\mu_{<\infty} + O\big(\Pi \log w_\mu^{-1}\big) + O\Big(\sqrt{\Pi(\log w_\mu^{-1})\mathbf{E}\ell^\mu_{<\infty}}\Big), \tag{15}$$

respectively.

The bound may seem weak to a reader familiar with another learning model, *prediction with expert advice*, which has received quite some attention since [LW89, Vov90]. Algorithms of this type are based on a class of experts rather than hypotheses, and proceed by randomly selecting experts according to a (non-Bayesian) posterior based on past performance of the experts. It is straightforward to use a hypothesis as an expert. Thus the experts theorems (for instance [HP05, Theorem 8(i)]) imply a bound similar to (15), but *without any assumption on the data generating process* μ, instead the bounds are relative to the best expert (hypothesis) in hindsight $\hat{\nu}$ (and moreover with $\log w_{\hat{\nu}}^{-1}\Pi(w)$ replaced by $\log w_{\hat{\nu}}^{-1}$). So the experts bounds are stronger, which does not necessarily imply that the experts algorithms are better: bounds like (15) are derived in the worst case over all loss functions, and in this worst case Bayesian learning is not better than experts learning, even under the proper learning assumption. However, experts algorithms do not provide estimates for the probabilities, which Bayesian algorithms do provide: in many practically relevant cases learning probabilities does yield superior performance.

2.5 Discussion

The proofs in this work 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. The potential method used here has been inspired by similar idea in prediction with expert advice [CBL03], the proof techniques are however completely different. We will in particular introduce the *entropy potential*, already stated in (11), which may be interpreted as the worst-case entropy of the model class under all admissible transformations of the weights, where the weight of the true distribution is kept fixed. The entropy potential is possibly a novel definition in this work.

Before starting the technical presentation, we discuss the limitations of our online learning setup. A Bayesian online learner defined in the straightforward way is computationally inefficient, if in each time step the full posterior is computed: Thus, marginalization, MAP/MDL, and stochastic model selection are equally inefficient in a naive implementation, and even generally uncomputable in case of a countable model class. On the other hand, many practical and efficient learning methods (e.g. training of an artificial neural network) are approximations to MAP/MDL and stochastic model selection. Moreover, bounds for the online algorithm also imply bounds for the *offline* variant, if additional assumptions (i.i.d.) on the process generating the inputs are satisfied. Also, in some cases one can sample efficiently from a probability distribution without knowing the complete distribution.

But the most important contribution of this paper is theoretical, as it clarifies the learning behavior of all three variants of Baysian learning in the ideal case. Also, countable hypothesis classes constitute the limit of what is computationally feasible at all, for this reason they are a core concept in Algorithmic Information Theory [LV97]. Proving corresponding results for the likewise important case of continuously parameterized model classes is, to our knowledge, an open problem.

As already indicated, the dependence of the bound (10) on w_μ^{-1} is logarithmic if the prior weights decay sufficiently rapidly (precisely polynomially), but linear in the worst case. This implies the practical recommendation of using a prior with light tails together with stochastic model selection.

The remainder of this paper is structured as follows. In the beginning of the next section, we will introduce the notation and, in order to introduce the methods, prove Solomonoff's result with a potential function. In Section 3.1, we consider stochastic model selection and prove the main auxiliary result. Section 3.2 defines the entropy potential and proves bounds for general countable model class. In Section 3.3 we turn to the question how large the newly defined entropy potential can be.

3 Technical results

The basic notation has been already introduced. We start with a simple example.

Example 6 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. $\mathcal{C}$ 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 $\Omega(-\log w_\mu)$ on the expected quadratic error of Bayes mixture and stochastic model selection predictions, namely the l.h.s. expressions of (8) and (23), respectively.

Next, we present a proof of Solomonoff's [Sol78] remarkable universal induction result, Theorem 1. The proof presented here slightly differs from the standard one [Hut04] and serves for introducing our main proof technique, namely potential functions.

Lemma 7 *Assume that the data generating distribution is contained in the model class, i.e.* $\mu \in \mathcal{C}$*. Define the* complexity potential *as*

$$\mathcal{K}(h_{<t}) = -\log w_\mu(h_{<t}). \tag{16}$$

For any current input z_t *and any history* $h_{<t}$*, this potential satisfies*

$$\begin{aligned} &(i) \quad \mathcal{K}(h_{<t}) \geq 0, \\ &(ii) \quad \mathcal{K}(h_{<t}) - \mathbf{E}_{x_t\sim\mu(\cdot|z_t)}\mathcal{K}(h_{1:t}) \geq \Delta_t^2(\mu,\xi). \end{aligned} \tag{17}$$

By summing up the expectation of (ii) while observing (i), this lemma immediately implies Theorem 1 for arbitrary sequence of inputs $z_1, z_2, \ldots$:

$$\sum_{t=1}^{\infty} \mathbf{E}_\mu \Delta_t^2(\mu,\xi) \leq \mathcal{K} = -\log w_\mu. \tag{18}$$

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. $\mathcal{K}$ now stands for $\mathcal{K}(h_{<t})$, please compare Remark 2 again),

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

The r.h.s. here is called *Kullback-Leibler divergence*. It is well-known that Kullback-Leibler divergence dominates the squared Hellinger distance $\Delta_t^2(\mu,\xi)$, see (20). □

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 $\mu(x_t) \geq \xi(x_t)$. 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 8 Consider a binary observation space 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 9.

3.1 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. (compare (6))

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

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 9 *If the current entropy of the hypothesis class is finite, $\mathcal{H}(h_{<t}) < \infty$, then, for any input z_t,*

$$\mathcal{H}(h_{<t}) - \mathbf{E}_{x_t \sim \xi(\cdot|z_t,h_{<t})}\mathcal{H}(h_{1:t}) \geq \mathbf{E}\Delta_t^2(\xi, \Xi). \tag{19}$$

Proof. It is a well-known fact, shown e.g. in [BM98, p. 178], that the squared Hellinger distance of two probability distributions μ and ν on $\mathcal{X}$ never exceeds their Kullback-Leibler divergence:

$$\Delta^2(\mu,\nu) = \int d\big(\sqrt{\mu(\cdot)} - \sqrt{\nu(\cdot)}\big)^2 \leq \int d\mu(\cdot) \log \frac{\mu(\cdot)}{\nu(\cdot)}. \tag{20}$$

Therefore, we have

$$\begin{aligned}\mathcal{H}(h_{<t}) - \mathbf{E}_{x_t\sim\xi(\cdot|z_t,h_{<t})}\mathcal{H}(h_{1:t}) &= \sum_{\nu\in\mathcal{C}} w_\nu(h_{<t}) \int d\nu(x|z_t)\log\tfrac{\nu(x|z_t)}{\xi(x|z_t,h_{<t})}\\ &\geq \sum_{\nu\in\mathcal{C}} w_\nu(h_{<t})\Delta^2(\xi,\nu) = \mathbf{E}\Delta_t^2(\xi,\Xi). \quad \Box\end{aligned}$$

□

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 10 (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,

$$\mathcal{P}_E(h_{<t}) - \mathbf{E}_{x_t\sim\mu(\cdot|z_t)}\mathcal{P}_E(h_{1:t}) \geq \Delta_t^2(\xi,\Xi). \tag{21}$$

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

$$\sum_{t=1}^{\infty}\Delta_t^2(\xi,\Xi) \leq \mathcal{P}_E = \mathcal{H}/w_\mu, \tag{22}$$

$$\sum_{t=1}^{\infty}\Delta_t^2(\mu,\Xi) \leq -\log(w_\mu) + \mathcal{H}/w_\mu + 2\sqrt{-\mathcal{H}\log(w_\mu)/w_\mu}, \tag{23}$$

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 9 we obtain (21) by

$$\begin{aligned}\mathcal{P}_E(h_{<t}) - \int d\mu(x|z_t)\mathcal{P}_E(h_{1:t}) &= \tfrac{1}{w_\mu(h_{<t})}\big(\mathcal{H}(h_{<t}) - \int d\xi(x|z_t,h_{<t})\mathcal{H}(h_{1:t})\big)\\ &\geq \Delta_t^2(\xi,\Xi).\end{aligned}$$

Summing the expectation up yields (22). Using this together with (18) and the triangle inequality for the Hellinger distance, we conclude (23). Finally, almost sure convergence follows from (12). □

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 11 Let the observation space 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, $E\Delta^2(\mu, \Xi) = \frac{1}{n}(2 - \sqrt{2}) > 1/(2n)$. Therefore the cumulative error exceeds $H_0/2$, 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 10, we used only one "wasteful" inequality, namely $\frac{1}{w_\mu(h_{<t})} \geq 1$. The following lemma will be our main tool for obtaining better bounds.

Lemma 12 (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:*

$$\begin{aligned} &(i) && B(h_{<t}) \geq \mathcal{H}(h_{<t}) \text{ (dominates the entropy),} \\ &(ii) && \mathbf{E}_{x_t \sim \mu(\cdot|z_t)} B(h_{1:t}) \leq B(h_{<t}) \text{ (decreases in expectation),} \\ &(iii) && \text{the value of } B(h_{<t}) \text{ can be approximated arbitrarily close} \\ &&& \text{by restricting to a finite model class.} \end{aligned}$$

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

$$\mathcal{P}(h_{<t}) - \mathbf{E}_{x_t \sim \mu(\cdot|z_t)} \mathcal{P}(h_{1:t}) \geq \mathcal{H}(h_{<t}) - \mathbf{E}_{x_t \sim \xi(\cdot|z_t, h_{<t})} \mathcal{H}(h_{1:t}). \tag{24}$$

Proof. The proof is divided into two parts. First, we will assume finite observation space. In the second part, this is generalized to arbitrary observation spaces.

Part 1: Assume that $\mathcal{X}$ is finite. 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

$$\begin{aligned}
&\mathcal{K} + \log(1+\mathcal{H}) - \sum_{x\in\mathcal{X}} \mu(x)\big[\mathcal{K}(x) + \log(1+\mathcal{H}(x))\big]\frac{1+B(x)}{1+B} \\
&\geq \mathcal{K} + \log(1+\mathcal{H}) - \sum_{x\in\mathcal{X}} \frac{\mu(x)(1+B(x))}{\sum_{x'}\mu(x')(1+B(x'))}\big[\mathcal{K}(x) + \log(1+\mathcal{H}(x))\big] \qquad (25)\\
&\geq \frac{\sum_\nu w_\nu \sum_x \nu(x)\log\frac{\nu(x)}{\xi(x)}}{1+B}, \qquad (26)
\end{aligned}$$

(25) follows from assumption (*ii*), so that we only need to show (26) in order to complete the proof. We will demonstrate an even stronger assertion:

$$\log(1+\mathcal{H}) - \sum_{x\in\mathcal{X}} \tilde{\mu}_x\big[\log(1+\mathcal{H}(x)) - \log\tfrac{\mu(x)}{\xi(x)}\big] \geq \frac{\sum_\nu w_\nu \sum_x \nu(x)\log\frac{\nu(x)}{\xi(x)}}{1+B} \qquad (27)$$

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 (27) 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 (27) follows from

$$f(\tilde{\mu}) = \log(1+\mathcal{H}) - \sum_{x\in\tilde{\mathcal{X}}} \tilde{\mu}_x\big(\tilde{V}(x) - \log\tfrac{\mu(x)}{\tilde{\xi}(x)}\big) \geq \frac{\sum_\nu w_\nu \sum_x \nu(x)\log\frac{\nu(x)}{\xi(x)}}{1+B}, \qquad (28)$$

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

$$\lambda = -\tilde{V}(x) + \log\tfrac{\mu(x)}{\tilde{\xi}(x)} \text{ for all } x \in \tilde{\mathcal{X}}. \qquad (29)$$

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 (29), (28) is transformed to

$$\begin{aligned}
\frac{\sum_{\nu\in\mathcal{C}} w_\nu \sum_{x\in\mathcal{X}} \nu(x)\log\frac{\nu(x)}{\xi(x)}}{1+B} \quad &\leq \quad \log(1+\mathcal{H}) + \lambda \qquad (30)\\
= \log(1 - \sum_{\nu\in\mathcal{C}} w_\nu \log w_\nu) &- \log\big[1 - \sum_{x\in\tilde{\mathcal{X}}}\tilde{\xi}(x)\sum_{\nu\in\mathcal{C}}\tfrac{w_\nu\nu(x)}{\tilde{\xi}(x)}\log\tfrac{w_\nu\nu(x)}{\xi(x)}\big].
\end{aligned}$$

The arguments of both outer logarithms on the r.h.s. of (30) 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 $\log(y)-\log(x) \geq \frac{y-x}{1+B}$, (30) 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.

Part 2: So far, we have proven the assertion for finite $\mathcal{X}$. In order to show the generalization to arbitrary observation space, we may decompose $\mathcal{X}$ into two subsets $\mathcal{X} = \mathcal{X}^{\text{discrete}} \cup \mathcal{X}^{\text{continuous}}$, where $\mathcal{X}^{\text{discrete}}$ is the at most countable set of points where any of the distributions in $\mathcal{C}$ has a non-zero mass concentration. We can prove the assertion for the discrete and the continuous parts separately. The discrete part follows simply by approximating, so we focus on the continuous part and assume without loss of generality that all distributions are (piecewise) continuous probability densities.

We show the assertion by assuming the contrary

$$\mathbf{E}_{x_t\sim\mu(\cdot|z_t)}\mathcal{P}(h_{1:t}) - \mathcal{P}(h_{<t}) > \mathbf{E}_{x_t\sim\xi}\mathcal{H}(h_{1:t}) - \mathcal{H}(h_{<t}) \tag{31}$$

and obtaining a contradiction to part 1. Dropping again the history $h_{<t}$ from the notation, (31) is equivalent to

$$\mathbf{E}_{x\sim\mu}\mathcal{P}(x) - \mathcal{P} > \mathbf{E}_{x\sim\xi}\mathcal{H}(x) - \mathcal{H} + 11\varepsilon \text{ for some } \varepsilon > 0. \tag{32}$$

We may assume without loss of generality that $\mathcal{X}$ is *compact*, and that there is a number $R > 0$ such that

$$\max_{x\in\mathcal{X}} \mathcal{H}(x)\xi(x) \leq R \text{ and } \max_{x\in\mathcal{X}} \xi(x) \leq R. \tag{33}$$

To see this, just choose $\tilde{\mathcal{X}} \subset \mathcal{X}$ compact and sufficiently large, such that both (33) and $\mathbf{E}_{x\sim\xi|_{\tilde{\mathcal{X}}}}\mathcal{H}(x) \geq \mathbf{E}_{x\sim\xi}\mathcal{H}(x) - \varepsilon$ hold, this is possible because $\xi(x)\mathcal{H}(x)$ is integrable w.r.t. the Lebesgue measure λ. Then, replace $\mathcal{X}$ by $\tilde{\mathcal{X}}$ and (32) by

$$\mathbf{E}_{x\sim\mu}\mathcal{P}(x) - \mathcal{P} > \mathbf{E}_{x\sim\xi}\mathcal{H}(x) - \mathcal{H} + 10\varepsilon. \tag{34}$$

Next, we argue that we may even assume without loss of generality that $\mathcal{C}$ is *finite*. To this aim, first start with approximating $\mathcal{P}(x)$ by a step function $\tilde{\mathcal{P}}(x)$ that is piecewise constant on relatively compact subsets $A_1, A_2, \ldots, A_n \subset \mathcal{X}$ and takes only finitely many (namely n) values $\tilde{y}_1, \ldots, \tilde{y}_n > 0$. We choose $\tilde{\mathcal{P}}(x)$ such that it is dominated by $\mathcal{P}(x)$, with the property

$$\mathbf{E}_{x\sim\mu}\tilde{\mathcal{P}}(x) \geq \mathbf{E}_{x\sim\mu}\mathcal{P}(x) - \varepsilon.$$

This is possible since $\mathcal{P}(x)$ is measurable and non-negative.

We choose an even smaller step function $\underline{\mathcal{P}}(x)$ that is likewise constant on $A_1, \ldots, A_n$ and is strictly dominated by $\tilde{\mathcal{P}}(x)$, such that

$$\mathbf{E}_{x\sim\mu}\underline{\mathcal{P}}(x) \geq \mathbf{E}_{x\sim\mu}\mathcal{P}(x) - 2\varepsilon \tag{35}$$

and $\underline{\mathcal{P}}(x) = \tilde{\mathcal{P}}(x) - \varepsilon_i = \tilde{y}_i - \varepsilon_i := y_i$ for $x \in A_i$, where $\varepsilon_i > 0$ for all $1 \leq i \leq n$. Then, for each $x \in A_1$, $B(x)$ and therefore $\mathcal{P}(x)$ can be approximated with finitely many hypotheses. Since $\mathcal{P}(x) \geq \tilde{y}_i$, we can find a finite set of hypotheses $\mathcal{F}(x)$ such that $\mathcal{P}^{\tilde{\mathcal{F}}}(x) > y_i$ for any $\tilde{\mathcal{F}} \supset \mathcal{F}(x)$, where $\mathcal{P}^{\tilde{\mathcal{F}}}(x)$ denotes the potential computed with only the hypotheses in $\tilde{\mathcal{F}}$. Since $\mathcal{P}^{\tilde{\mathcal{F}}}(\cdot)$ is continuous (while $\tilde{\mathcal{F}}$ is fixed), we have that $\mathcal{P}^{\tilde{\mathcal{F}}}(\tilde{x}) > y_i$ holds even within an open superset $\tilde{x} \in \mathcal{U}(x)$ of x. For each $x \in A_1$, there is such an open $\mathcal{U}(x)$, and they form an open cover of $\bar{A}_1$. Since $\bar{A}_1$ is compact, there is a finite subcover $\mathcal{U}(x_1) \cup \ldots \cup \mathcal{U}(x_m) \supset A_1$. We may choose $\mathcal{F}_1 = \mathcal{F}(x_1) \cup \ldots \cup \mathcal{F}(x_m)$ in order to obtain a finite set of hypotheses approximating $\mathcal{P}(x)$ sufficiently closely on all of A_1.

Analogous approximations $\mathcal{F}_2, \ldots, \mathcal{F}_n$ are obtained for all other $A_2, \ldots, A_n$. Also, we choose a finite set of hypotheses $\mathcal{F}_0 \subset \mathcal{C}$ such that all supersets $\tilde{\mathcal{F}} \supset \mathcal{F}_0$ approximate the prior $\mathcal{P}$ up to ε. Take the union $\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1 \cup \ldots \cup \mathcal{F}_n$. Then, from (35), we conclude that

$$\mathbf{E}_{x\sim\mu}\underline{\mathcal{P}}(x) - \mathcal{P}^{\tilde{\mathcal{F}}} > \mathbf{E}_{x\sim\mu}\mathcal{P}(x) - \mathcal{P} - 3\varepsilon.$$

and $\mathcal{P}^{\tilde{\mathcal{F}}}(x) \geq \underline{\mathcal{P}}(x)$ for all $x \in \mathcal{X}$ and any $\tilde{\mathcal{F}} \supset \mathcal{F}$. We make sure that $\mu \in \mathcal{F}$.

We perform the same construction an approximation of $\mathcal{H}(x)$ from above. Since $\mathcal{X}$ was already assumed to be compact, the constant function R, which dominates $\xi(x)\mathcal{H}(x)$ according to (33) is integrable w.r.t. the Lebesgue measure λ. Therefore, we may refine the partitioning $(A_i)_{i=1}^k$ of $\tilde{X}$, obtaining a new partitioning $(\tilde{A}_i)_{i=1}^m$ of $\tilde{X}$, such that $\mathcal{H}(x)\xi(x)$ is approximated from above within ε by functions constant on each $\tilde{A}_i$. We may choose the approximators $\overline{\mathcal{H}}(x)$ and $\overline{\xi}(x)$ slightly larger, such that they need only finitely many hypotheses. We incorporate these hypotheses into $\mathcal{F}$.

Altogether, this shows that we may indeed assume that $\mathcal{C}$ is finite, if we replace (34) with

$$\mathbf{E}_{x\sim\mu}\underline{\mathcal{P}}(x) - \mathcal{P} > \mathbf{E}_{x\sim\overline{\xi}}\overline{\mathcal{H}}(x) - \mathcal{H} + 4\varepsilon, \tag{36}$$

knowing that $\underline{\mathcal{P}}(x) \leq \mathcal{P}(x)$ and $\overline{\mathcal{H}}(x) \geq \mathcal{H}(x)$ for all $x \in \mathcal{X}$.

In the next step, we further decrease $\mathcal{X}$ a tiny little bit and define $\underline{\mathcal{X}} \subset \mathcal{X}$ such that

$$\nu(\underline{\mathcal{X}}) < 1 \text{ for all (finitely many!) } \nu \in \mathcal{C}. \tag{37}$$

Set $\underline{A}_i = A_i \cap \underline{\mathcal{X}}$ for all $1 \leq i \leq n$. While choosing $\underline{\mathcal{X}}$, we make sure that it is not too small. Namely, we assert

$$\mu(A_i)\Big(1 - \frac{\varepsilon}{2y_i}\Big) < \mu(\underline{A}_i) \text{ for all } 1 \leq i \leq n, \tag{38}$$

$$\mathbf{E}_{x\sim\mu|\underline{\mathcal{X}}}\underline{\mathcal{P}}(x) \geq \mathbf{E}_{x\sim\mu}\underline{\mathcal{P}}(x) - \varepsilon, \text{ and} \tag{39}$$

$$1 - \xi(\underline{\mathcal{X}}) < \frac{\varepsilon}{2\log|\mathcal{C}|}, \tag{40}$$

where $|\mathcal{C}|$ is the number of hypotheses in $\mathcal{C}$.

In the last step, construct a refining partition $(A'_i)_{i=1}^k$ of $(\tilde{A}_i \cap \underline{\mathcal{X}})_{i=1}^m$ and lower and upper approximations $\underline{\nu}, \overline{\nu}$ for each $\nu \in \mathcal{C}$, with the following properties:

$$\int_{\underline{X}} \overline{\nu} < 1 \text{ for all } \nu \in \mathcal{C}, \text{ possible due to (37),} \tag{41}$$

$$\int_{\underline{A}_i} \underline{\mu} \geq \mu(\underline{A}_i)\Big(1 - \frac{\varepsilon}{y_i}\Big) \text{ for all } 1 \leq i \leq n, \text{ possible due to (38),} \tag{42}$$

$$1 - \int_{\underline{X}} \underline{\xi} < \frac{\varepsilon}{\log|\mathcal{C}|}, \text{ possible due to (40).} \tag{43}$$

Now choose (with λ being the Lebesgue measure)

$$\begin{aligned} &\mathcal{X}' = \{0, 1, \ldots, k\}, && x'_i = \arg\min_{x \in A'_i} \mathcal{P}(x) \text{ for all } 1 \leq i \leq k, \\ &\nu'_i = \nu(x'_i)\lambda(A'_i) \quad (1 \leq i \leq k, \nu \in \mathcal{C}), && \nu'_0 = 1 - \sum_{i=1}^k \nu'_i \text{ for all } \nu \in \mathcal{C}, \\ &\mathcal{H}'_i = \mathcal{H}\left(\Big(\frac{w_\nu \nu'_i}{\sum_{j=0}^k w_\nu \nu'_j}\Big)_{\nu \in \mathcal{C}}\right) = \mathcal{H}(x'_i), && \mathcal{P}'_i = \mathcal{P}\left(\Big(\frac{w_\nu \nu'_i}{\sum_{j=0}^k w_\nu \nu'_j}\Big)_{\nu \in \mathcal{C}}\right) = \mathcal{P}(x'_i). \end{aligned}$$

By (41), each ν' is in fact a measure on $\mathcal{X}'$. Justifying the following estimations with the respective equations before, we have

$$\begin{aligned} \sum_{i=0}^k \mu'_i \mathcal{P}'_i - \mathcal{P} \geq \sum_{i=1}^k \mu'_i \mathcal{P}'_i - \mathcal{P} \quad &\geq \mathbf{E}_{\underline{\mu}|\underline{X}} \underline{\mathcal{P}} - \mathcal{P} \quad \overset{(42)}{\geq} \mathbf{E}_{\mu|\underline{X}} \underline{\mathcal{P}} - \mathcal{P} - \varepsilon \\ \overset{(39)}{\geq} \mathbf{E}_\mu \underline{\mathcal{P}} - \mathcal{P} - 2\varepsilon \quad &\overset{(36)}{\geq} \mathbf{E}_{\overline{\xi}} \overline{\mathcal{H}} - \mathcal{H} + 2\varepsilon \quad \geq \mathbf{E}_{\overline{\xi}|\underline{X}} \overline{\mathcal{H}} - \mathcal{H} + 2\varepsilon \\ \geq \sum_{i=1}^k \xi'_i \mathcal{H}'_i - \mathcal{H} + 2\varepsilon \overset{(43)}{\geq} \sum_{i=0}^k \xi'_i \mathcal{H}'_i - \mathcal{H} + \varepsilon. \end{aligned}$$

The last estimate is true since $\mathcal{H}'_0 \leq \log|\mathcal{C}|$ holds. This is the desired contradiction to the first part of the proof. □

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

Theorem 13 (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,

$$\sum_{t=1}^\infty \Delta^2(\xi, \Xi) \leq \mathcal{P}_F = (1 + \log N)\big[\log(1 + \mathcal{H}) - \log(w_\mu)\big]. \tag{44}$$

Proof. Since the entropy of a class with N elements is at most $\log N$, this follows directly from Lemma 12. □

3.2 Entropy potential and countable classes

We now generalize Theorem 13 to arbitrary countable model classes. First note that there is one very convenient fact about the potential function proofs so far: (17), (21), and (24) 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 12, 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 14 Fix binary observation space and let $\tilde{\mathcal{C}}$ and $\tilde{\mathcal{Z}}$ be model class and input space of Example 11. 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 11. 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) $H_0/2$, while the entropy can be made arbitrarily small for large m.

Definition 15 *(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, as already stated in (11),

$$\Pi\big((w_\nu)_{\nu\in\mathcal{C}}\big) = \sup\big\{H\big((\tfrac{\tilde{w}_\nu}{\sum_{\nu'}\tilde{w}_{\nu'}})_\nu\big) : \tilde{w}_\mu = w_\mu \wedge \tilde{w}_\nu \le w_\nu \ \forall \nu \in \mathcal{C}\setminus\{\mu\}\big\}. \tag{45}$$

Clearly, $\Pi \ge \mathcal{H}$. According to Theorem 10, 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 12.

The following proposition gives a characterization of the entropy potential.

Proposition 16 (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*

$$-\log w_\nu > L(A) := -\sum_{\nu'\in A} \tfrac{w_{\nu'}}{w(A)} \log w_{\nu'} \quad \Longleftrightarrow \quad \nu \in A\setminus\{\mu\}. \tag{46}$$

We call A the set of active *models (in Π). Then, with $\tilde{w}_\nu = \exp(-L(A))$ for $\nu \in \mathcal{C}\setminus A$, $\tilde{w}_\nu = w_\nu$ for $\nu \in A$, and $k = |\mathcal{C}\setminus A|$, we have*

$$\begin{aligned}\Pi = \Pi\big((w_\nu)_{\nu\in\mathcal{C}}\big) &= H\big((\tfrac{\tilde{w}_\nu}{\sum_{\nu'}\tilde{w}_{\nu'}})_{\nu\in\mathcal{C}}\big) \\ &= \log\big(k + w(A)e^{L(A)}\big).\end{aligned} \tag{47}$$

Moreover, this is scaling invariant in the weights, i.e. (46) yields the correct active set and (47) 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 (45) cannot be attained if some $\tilde{w}_\nu = 0$. To this aim, let $\tilde{p} \in [0,1]^{|\mathcal{C}|}$, assume $\tilde{w}_\nu = 0$ for a specific $\nu \in \mathcal{C}$, and set $\tilde{H} = H\big((\frac{\tilde{w}_\nu}{\sum_{\nu'}\tilde{w}_{\nu'}})_\nu\big)$. Now assume $\tilde{w}_\nu > 0$ and observe that

$$\begin{aligned}&H\big((\tfrac{\tilde{w}_\nu}{\sum_{\nu'}\tilde{w}_{\nu'}})_\nu\big) = \\ &\quad -\tilde{w}_\nu \log(\tilde{w}_\nu) + (1-\tilde{w}_\nu)\big[-\log(1-\tilde{w}_\nu) + \tilde{H}\big] \geq \tilde{H}\end{aligned}$$

holds if $-\log(\tilde{w}_\nu) \geq \tilde{H}$. This can be realized for small enough $\tilde{w}_\nu > 0$, hence the maximum of (45) cannot be attained for $\tilde{w}_\nu = 0$.

The maxima of the entropy under $\tilde{p}$ can be found only at the boundary or at the points where the gradient vanishes. Therefore, for a maximum of (45), we need that for each $\nu \in \mathcal{C} \setminus \{\mu\}$, either $\tilde{w}_\nu = w_\nu$ or, with $\tilde{p}(\mathcal{C}) = \sum_\nu \tilde{w}_\nu$ and $\tilde{L}(\mathcal{C}) = -\frac{1}{\tilde{p}(\mathcal{C})}\sum_\nu \tilde{w}_\nu \log(\tilde{w}_\nu)$,

$$0 = \frac{dH\big((\frac{\tilde{w}_\nu}{\sum_{\nu'}\tilde{w}_{\nu'}})_\nu\big)}{d\tilde{w}_\nu} = \frac{1}{\tilde{p}(\mathcal{C})}\Big[-\log(\tilde{w}_\nu) - \tilde{L}(\mathcal{C})\Big]. \tag{48}$$

Those elements ν satisfying the latter condition have $\log \tilde{w}_\nu = -\tilde{L}(\mathcal{C})$ and hence $\tilde{L}(\mathcal{C}) = \tilde{L}(\mathcal{C}\setminus\{\nu\})$. Therefore, each possible maximum of (45) corresponds to a subset $\tilde{A} \subset \mathcal{C}$ of *active models*, such that $\mu \in \tilde{A}$ and furthermore $\tilde{w}_\nu = w_\nu$ for $\nu \in \tilde{A}$ and $\log \tilde{w}_\nu = -\tilde{L}(\tilde{A}) = -L(\tilde{A})$ for $\nu \notin \tilde{A}$. Since only $\log \tilde{w}_\nu \leq \log w_\nu$ 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 $-\log \tilde{w}_\nu$ equals the average complexity of all $\nu \in \tilde{A}$. Hence

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

which proves (47) 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 (48), is non-positive at $\tilde{w}_{\nu_1} = w_{\nu_1}$. Thus we may increase the entropy by decreasing

w_{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 (46) and the value (47) w.r.t. the weights is easy to see. □

The next result states that the entropy potential possesses the desired property to *decrease in expectation* and therefore paves the way for the main theorem of this work.

Theorem 17 *For any history $h_{<t}$ and current input z_t,*

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

where the posterior entropy potential is defined as $\Pi(h_{<t}) := \Pi\big([w_\nu(h_{<t})]_{\nu\in\mathcal{C}}\big)$.

Proof. As in the proof of Lemma 12, we need to proceed in two steps. First we show the assertion for finite observation space $\mathcal{X}$. The second step, the generalization to arbitrary $\mathcal{X}$, is similar as but substantially simpler that the second part of the proof of Lemma 12, since the approximation of only one side of the bound is required. We therefore omit the explicit presentation of this second step here.

Restricting to finite observation space $\mathcal{X}$, 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

$$\sum_{x\in\tilde{\mathcal{X}}} \mu(x)\Pi(x) \leq \mu(\tilde{\mathcal{X}})\Pi\Big(\big[w_\nu\nu(\tilde{\mathcal{X}})\big]_{\nu\in\mathcal{C}}\Big), \tag{49}$$

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 (49) by induction over the subset size $|\tilde{\mathcal{X}}|$. For $|\tilde{\mathcal{X}}| = 1$, there is nothing to show. If (49) 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

$$(1-\tilde{\mu})\Pi\big([\tilde{w}_\nu(1-p_\nu)]_{\nu\in\mathcal{C}}\big) + \tilde{\mu}\Pi\big([\tilde{w}_\nu p_\nu]_{\nu\in\mathcal{C}}\big) \leq \Pi\big([\tilde{w}_\nu]_{\nu\in\mathcal{C}}\big), \tag{50}$$

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 (50), 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 16). Denote the respective sets of active models by A, A^0, A^1. Recall that the constructions in Proposition 16 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 $p_{\nu'}$ ($\nu' \neq \nu$),

$$\text{there is an assignment } p_\nu \in [0,1] \text{ such that both } \nu \in A^0 \text{ and } \nu \in A^1 \tag{51}$$

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 (51). 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 $\tilde{w}_\nu < e^{-L(A)}$,

$$\frac{1}{e^{\sum_A \frac{\tilde{w}_\nu(1-p\nu)}{\tilde{w}^0(A)} \log \frac{1}{\tilde{w}_\nu(1-p\nu)}}} + \frac{1}{e^{\sum_A \frac{\tilde{w}_\nu p\nu}{\tilde{w}^1(A)} \log \frac{1}{\tilde{w}_\nu p\nu}}} = e^{-L^0(A)} + e^{-L^1(A)} \geq e^{-L(A)} \tag{52}$$

implies that I is not empty. We will verify (52) 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 (52), observe that

$$e^{\sum_A \frac{\tilde{w}_\nu (1-p\nu)}{\tilde{w}^0(A)} \log \frac{1}{(1-p\nu)}} \leq \frac{\tilde{w}(A)}{\tilde{w}^0(A)} \quad \text{and} \quad e^{\sum_A \frac{\tilde{w}_\nu p\nu}{\tilde{w}^1(A)} \log \frac{1}{p\nu}} \leq \frac{\tilde{w}(A)}{\tilde{w}^1(A)}$$

by Jensen's inequality, so (52) follows from

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

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 (50), 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 18 below, and so is the Hessian of Π^0. This concludes the proof. □

Lemma 18 *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*

$$\sum_{i,j=1}^N u_i u_j \left[\frac{\mathbb{1}_{i=j} W}{w_i} - \log w_i - L - \log w_j - L - 1\right] \geq 0. \tag{53}$$

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 (53) 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 (53) by decreasing w_0, until eventually $-\log w_k = L$ holds for one $k \in \{1 \ldots N\}$. Set $D_i = -\log w_i - L$ and $\tilde{W} = W - w_k$. Then

$$\sum_{i,j=1}^N u_i u_j \left[\frac{\mathbb{1}_{i=j} W}{w_i} + D_i + D_j - 1\right] = \sum_{i,j \in \{1\ldots N\}\setminus\{k\}} u_i u_j \left[\frac{\mathbb{1}_{i=j} \tilde{W}}{w_i} + D_i + D_j - 1\right]$$
$$+ \sum_{i \in \{1\ldots N\}\setminus\{k\}} \left[\frac{w_i}{w_k} u_k^2 - 2(1 - D_i) u_i u_k + \frac{w_k}{w_i} u_i^2\right]. \tag{54}$$

Since for all $u, v \in \mathbb{R}$ and $c \leq 1$ we have $u^2 - 2cuv + v^2 \geq 0$, the term (54) is nonnegative. Thus the assertion follows from the induction hypothesis. □

The previous theorem, together with Lemma 12, immediately implies the main result of this paper, Theorem 4. More precisely, it reads as follows.

Theorem 19 (Predictive performance of stochastic model selection) *For countable model class $\mathcal{C}$ containing the true distribution μ, define the potential (for errors in the squared Hellinger sense) 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,

$$\mathcal{P}(h_{<t}) - \mathbf{E}_{x_t \sim \mu(\cdot|z_t)} \mathcal{P}(h_{1:t}) \geq \mathcal{H}(h_{<t}) - \mathbf{E}_{x_t \sim \xi(\cdot|z_t, g_{<t})} \mathcal{H}(h_{1:t}), \textit{ and thus}$$
$$\sum_{t=1}^\infty \Delta_t^2(\xi, \Xi) \leq \mathcal{P} = (1 + \Pi)\big[\log(1 + \mathcal{H}) - \log(w_\mu)\big]. \tag{55}$$

The bound (10) on $\sum_t \Delta_t^2(\mu, \Xi)$ follows from the triangle inequality for the Hellinger distance.

3.3 The magnitude of the entropy potential

In this section, we will answer the question how large the newly defined quantity, the entropy potential Π, can grow.

Theorem 20 *Let A denote the active set from the Proposition 16. The entropy potential Π of a discrete distribution $(w_\nu)_{\nu\in\mathcal{C}}$ has the following properties.*

(i) It is lower bounded by the entropy of the distribution, $\Pi \geq H\big((w_\nu)_{\nu\in\mathcal{C}}\big)$.

(ii) It is finite if and only if the entropy is finite, hence, in the definition the supremum can be replaced by a maximum.

(iii) The upper bound $\Pi \leq H\big((w_\nu)_{\nu\in\mathcal{C}}\big)w_\mu^{-1}$ always holds.

(iv) There are cases where this bound is sharp within a multiplicative constant.

(v) $\Pi \leq L(A)$, where $L(A)$ from Proposition 16 is used.

(vi) With $\nu_1 := \arg\max\big\{w_\nu : \nu \in A\backslash\{\mu\}\big\}$ and A from Proposition 16, $\Pi \leq -\log w_{\nu_1}$ holds.

Proof. Part (i) is obvious. The same holds for (v), this can be seen by observing $k \leq \frac{1-p(A)}{e^{-L(A)}}$ and (47). Part (ii) follows from (iii), which can be verified by

$$\mathcal{H} \geq -\sum_{\nu\in A} w_\nu \log w_\nu = p(A)L(A) \geq w_\mu L(A) \geq w_\mu \Pi,$$

where (v) was used. Since $L(A) \leq -\log w_{\nu_1}$, (v) also implies (vi).

It remains to show (iv). To this aim, fix large numbers $n, m \in \mathbb{N}$ and set $K = 2^{nm}$. Let

$$\mathcal{C} = \{\mu, \nu_{\text{fool}}, \nu_1, \nu_2, \ldots, \nu_K\}.$$

Set the probability distribution to

$$w_{\text{fool}} = 1 - \frac{1}{m},\ w_\mu = \frac{1}{m}(1 - \frac{1}{n}),\ \text{and } p_i = \tfrac{1}{mnK}$$

for $1 \leq i \leq K$. Then for the entropy of the model class, $H\big((p)\big) \approx \log 2$, while $w_\mu \approx \frac{1}{m}$ and $\Pi \approx m\log 2 \approx \mathcal{H}/w_\mu$ holds. □

The general bound in Theorem 20 (iii) does not exclude cases where the entropy potential is huge. And in fact, there are cases where this bound is sharp up to a factor, and also the cumulative quadratic Hellinger distance is of the same order:

$$\sum_{t=1}^{\infty} \Delta^2(\xi, \Xi) = \Omega\big(\Pi\big) = \Omega\big(\tfrac{\mathcal{H}}{w_\mu}\big). \tag{56}$$

In order to see this, consider the case of Example 14 and choose large $m, n > 1$ and $H_0 := m$. Then $\mathcal{H} \approx \log 2$, $w_\mu \approx \frac{1}{m}$, and $\Pi \approx H_0 \log 2 \approx \mathcal{H}/w_\mu$. Moreover, as seen above, the expected cumulative quadratic Hellinger error exceeds $\frac{1}{2}H_0$. Hence, for this model class and prior, (56) holds.

Fortunately, for broad classes of discrete distributions, it is of reasonable size $O(-\log w_\mu)$. Precisely, this happens if the *tails* of the probability distribution p are sufficiently light.

Proposition 21 (*i*) *If w_ν decays exponentially, $\Pi = O(-\log w_\mu)$ holds. For simplicity, we may identify ν with its index in an enumeration, then exponential decay is reads as $w_\nu = O(\alpha^\nu)$ for some $\alpha \in (0,1)$.*
(*ii*) *If w_ν decays inverse polynomially, that is, $w_\nu = O(\nu^{-b})$ for $b > 1$, we have $\Pi = O\big(-\frac{b^2}{b-1}\log w_\mu\big)$.*

Note that the entropy potential does not depend on a reordering of the elements of $\mathcal{C}$. That is, we just need that there exists some reordering of $\mathcal{C}$ such that the probabilities decay in the stated way.

Proof. We simplify the exposition and assume that the weights decay *exactly* exponentially or polynomially, respectively. The general case follows easily.

Therefore, in order to show (*i*), we identify $\mathcal{C}$ with $\{0, 1, \ldots\}$ and assume that $p_i = \alpha^i(1-\alpha)$, for $i = 0, 1, \ldots$ and $\alpha \in (0,1)$. For given $k \cong \mu$, i.e., $k \geq 0$ is the index of the reference element μ, we need to find an index $j \geq 0$ such that

$$-j\log\alpha \geq -\frac{\alpha^k k\log\alpha + \sum_{i=j}^\infty \alpha^i i\log\alpha}{\alpha^k + \sum_{i=j}^\infty \alpha^i}. \tag{57}$$

Then, we know by (46) that the active set is a superset of $\{j, j+1, \ldots\}$, and in particular $\Pi \leq -\log p_j = O(j)$ according to Theorem 20 (*vi*).

By some elementary transformations, we see that (57) holds if $j(\alpha^k + \frac{\alpha^j}{1-\alpha}) \geq k\alpha^k + \frac{\alpha^{j+1}}{(1-\alpha)^2} + \frac{j\alpha^j}{1-\alpha}$. This follows from $j \geq k + \frac{\alpha^2}{(1-\alpha)^2}$, which is therefore sufficient to state one index j which is necessarily contained in the active set. Hence $O(j) = O(k)$, and (*ii*) is proven. (We remark that Theorem 20 (*vi*) gives a rough estimate here, in reality the order of the entropy potential is even smaller.)

Part (*ii*) is shown in a similar way. Again we identify $k \cong \mu$. Here we assume $p_i \simeq i^{-b}$ for $i \geq 1$ and $b > 1$. Then we need to find an index $j \geq 1$ satisfying

$$\log j \geq \frac{k^{-b}\log k + \sum_{i=j}^\infty i^{-b}\log i}{k^{-b} + \sum_{i=j}^\infty i^{-b}}. \tag{58}$$

Since (46) is scale invariant, this gives the desired subset of the active set. With some transformations, we see that $k^{-b}\big(\log j - \log k\big) \geq \sum_{i=j}^\infty i^{-b}\big(\log i - \log j\big)$ implies (58). We may upper bound the sum on the r.h.s. by an integral and search for j such that

$$\begin{aligned} k^{-b}\big(\log j - \log k\big) &\geq 1 + \int_j^\infty x^{-b}\big(\log x - \log j\big)dx \\ &= 1 + \tfrac{1}{(b-1)^2} j^{-(b-1)} \end{aligned}$$

(the additional 1 stems from estimating the sum by the integral). Setting $j = k^{\frac{b}{b-1}}$ satisfies this requirement for sufficiently large k and thus completes the proof of (ii). □

If however the probability distribution decays very slowly, the entropy potential is of exponential order.

Theorem 22 *If w_ν decays as $\nu^{-1}(\log\nu)^{-b}$ for $b > 2$ and $\nu = 1, 2, \ldots$, we have $\Pi = \Omega(w_\mu^{-\frac{1}{b+1}})$.*

Proof. As in the proof of the last result, we assume that exactly $p_i \simeq i^{-1}(\log i)^-b$ holds for $i = 1, 2, \ldots$ and $b > 2$. We now need to show that for small j,

$$L_j < \frac{k^{-1}(\log k)^{-b}L_k + \sum_{i=j}^{\infty} i^{-1}(\log i)^{-b}L_i}{k^{-1}(\log k)^{-b} + \sum_{i=j}^{\infty} i^{-1}(\log i)^{-b}},$$

where $L_i = \log i + b\log\log i$. This is satisfied if

$$k^{-1}(\log k)^{-b}\big[\log j - \log k + b(\log\log j - \log\log k\big] < \\ \sum_{i=j}^{\infty} i^{-1}(\log i)^{-b}\big[\log i - \log j + b(\log\log i - \log\log j\big] \tag{59}$$

holds. The r.h.s. sum can be approximated by an integral which evaluates to

$$\frac{1}{(b-1)(b-2)(\log j)^{b-1}} + \frac{b}{(b-1)^2(\log j)^{b-2}}. \tag{60}$$

We now set $j = \lceil\exp(k^{\frac{1}{b}})\rceil$. Then for sufficiently large k, the first term of (60) dominates $k^{-1}(\log k)^{-b}\log j$. Moreover, for sufficiently large k, the second term of (60) dominates $k^{-1}(\log k)^{-b}\log\log j$. Hence, this choice of j satisfies (59) for sufficiently large k.

Given that the smallest index in $A \setminus \{\mu\}$ exceeds $j = \lceil\exp(k^{\frac{1}{b}})\rceil$, how large is Π at least? The answer is: $\Pi \geq \frac{1}{2}L_j = \frac{1}{2}(\log j + b\log\log j) \geq \frac{1}{2}k^{\frac{1}{b}}$, for k sufficiently large. This can be seen easily, we just need to make sure that the contribution of the tail $A \setminus \{\mu\}$ exceeds $\frac{1}{2}$. Approximating the tail weight by an integral, $\int_j^\infty \frac{dx}{x(\log x)^b} = \frac{1}{(b-1)(\log j)^{b-1}}$, this contribution is $\frac{1}{b-1}k^{-\frac{b-1}{b}}$ and therefore exceeds the contribution of the element μ, namely $k^{-1}(\log k)^{-b}$, for large k.

Finally, the proof is concluded by observing $\Pi = \Omega(k^{\frac{1}{b}}) = \Omega(w_\mu^{-\frac{1}{b+1}})$. □

The entropy potential is infinite with the usual definition of a *universal model class* [LV97]. But with a slight modification of the prior, it becomes finite. Hence we can obtain a universal induction result for stochastic model selection:

Example 23 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 19 implies consistency of universal stochastic model selection with this prior and normalization. Had we 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 growth must necessarily result in an infinite sum. Hence the bound for universal stochastic model selection is infinite with the usual prior.

References

[BC91] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, 37(4):1034–1054, 1991.

[BD62] D. Blackwell and L. Dubins. Merging of opinions with increasing information. *Annals of Mathematical Statistics*, 33:882–887, 1962.

[BM98] A. A. Borovkov and A. Moullagaliev. *Mathematical Statistics.* Gordon & Breach, 1998.

[BRY98] A. R. Barron, J. J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory*, 44(6):2743–2760, 1998.

[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.

[CD05] J. W. Comley and D. L. Dowe. Minimum message length and generalized Bayesian nets with asymmetric languages. In P. Grünwald, I. J. Myung, and M. A. Pitt, editors, *Advances in Minimum Description Length: Theory and Applications*, pages 265–294. 2005.

[GL04] P. Grünwald and J. Langford. Suboptimal behaviour of Bayes and MDL in classification under misspecification. In *17th Annual Conference on Learning Theory (COLT)*, pages 331–347, 2004.

[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.

[LW89] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. In *30th Annual Symposium on Foundations of Computer Science*, pages 256–261, Research Triangle Park, North Carolina, 1989. IEEE.

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

[PH06] J. Poland and M. Hutter. MDL convergence speed for Bernoulli sequences. *Statistics and Computing*, 16:161–175, 2006.

[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.

[Vov90] V. G. Vovk. Aggregating strategies. In *Proc. Third Annual Workshop on Computational Learning Theory*, pages 371–383, Rochester, New York, 1990. ACM Press.

[WB68] C. S. Wallace and D. M. Boulton. An information measure for classification. *Computer Jrnl.*, 11(2):185–194, August 1968.

[WD99] C. S. Wallace and D. L. Dowe. Minimum Message Length and Kolmogorov Complexity. *Computer Journal*, 42(4):270–283, 1999.