On Finding Predictors for Arbitrary Families of Processes

Daniil Ryabko

On Finding Predictors for Arbitrary Families of Processes

Daniil Ryabko
daniil@ryabko.net, INRIA Lille

Abstract

The problem is sequence prediction in the following setting. A sequence $x_1, \dots, x_n, \dots$ of discrete-valued observations is generated according to some unknown probabilistic law (measure) μ. After observing each outcome, it is required to give the conditional probabilities of the next observation. The measure μ belongs to an arbitrary but known class $\mathcal{C}$ of stochastic process measures. We are interested in predictors ρ whose conditional probabilities converge (in some sense) to the "true" μ-conditional probabilities if any $\mu \in \mathcal{C}$ is chosen to generate the sequence. The contribution of this work is in characterizing the families $\mathcal{C}$ for which such predictors exist, and in providing a specific and simple form in which to look for a solution. We show that if any predictor works, then there exists a Bayesian predictor, whose prior is discrete, and which works too. We also find several sufficient and necessary conditions for the existence of a predictor, in terms of topological characterizations of the family $\mathcal{C}$, as well as in terms of local behaviour of the measures in $\mathcal{C}$, which in some cases lead to procedures for constructing such predictors.

It should be emphasized that the framework is completely general: the stochastic processes considered are not required to be i.i.d., stationary, or to belong to any parametric or countable family.

1 Introduction

Given a sequence $x_1, \dots, x_n$ of observations $x_i \in \mathcal{X}$, where $\mathcal{X}$ is a finite set, we want to predict what are the probabilities of observing $x_{n+1} = x$ for each $x \in \mathcal{X}$, or, more generally, probabilities of observing different $x_{n+1}, \dots, x_{n+h}$, before x_{n+1} is revealed, after which the process continues. It is assumed that the sequence is generated by some unknown stochastic process μ, a probability measure on the space of one-way infinite sequences $\mathcal{X}^\infty$. The goal is to have a predictor whose predicted probabilities converge (in a certain sense) to the correct ones (that is, to μ-conditional probabilities). In general this goal is impossible to achieve if nothing is known about the measure μ generating the sequence. In other words, one cannot have a predictor whose error goes to zero for any measure μ. The problem becomes tractable if we assume that the measure μ generating the data belongs to some known class $\mathcal{C}$. The questions addressed in this work are a part of the following general problem: given an arbitrary set $\mathcal{C}$ of measures, how can we find a predictor that performs well when the data is generated by any $\mu \in \mathcal{C}$, and whether it is possible to find such a predictor at all. An example of a generic property of a class $\mathcal{C}$ that allows for construction of a predictor, is that $\mathcal{C}$ is countable. Clearly, this condition is very strong. An example, important from the applications point of view, of a class $\mathcal{C}$ of measures for which predictors are known, is the class of all stationary measures. The general question, however, is very far from being answered.

The contribution of this work to solving this question is, first, in that we provide a specific form in which to look for a predictor. More precisely, we show that if a predictor that predicts

every $\mu \in \mathcal{C}$ exists, then such a predictor can also be obtained as a weighted sum of countably many elements of $\mathcal{C}$. This result can also be viewed as a justification of the Bayesian approach to sequence prediction: if there exists a predictor which predicts well every measure in the class, then there exists a Bayesian predictor (with a rather simple prior) that has this property too. In this respect it is important to note that the result obtained about such a Bayesian predictor is pointwise (holds for every μ in $\mathcal{C}$), and stretches far beyond the set its prior is concentrated on. Next, we derive some characterizations of families $\mathcal{C}$ for which a predictor exist. We first analyze what is furnished by the notion of separability, when a suitable topology can be found: we find that it is a sufficient but not always a necessary condition. We then derive some sufficient conditions for the existence of a predictor which are based on local (truncated to the first n observation) behaviour of measures in the class $\mathcal{C}$. Necessary conditions cannot be obtained in this way (as we demonstrate), but sufficient conditions, along with rates of convergence and construction of predictors, can be found.

The **motivation** for studying predictors for arbitrary classes $\mathcal{C}$ of processes is two-fold. First of all, prediction is a basic ingredient for constructing intelligent systems. Indeed, in order to be able to find optimal behaviour in an unknown environment, an intelligent agent must be able, at the very least, to predict how the environment is going to behave (or, to be more precise, how relevant parts of the environment are going to behave). Since the response of the environment may in general depend on the actions of the agent, this response is necessarily non-stationary for explorative agents. Therefore, one cannot readily use prediction methods developed for stationary environments, but rather has to find predictors for the classes of processes that can appear as a possible response of the environment.

Apart from this, the problem of prediction itself has numerous applications in such diverse fields as data compression, market analysis, bioinformatics, and many others. It seems clear that prediction methods constructed for one application cannot be expected to be optimal when applied to another. Therefore, an important question is how to develop specific prediction algorithms for each of the domains.

Prior work. As it was mentioned, if the class $\mathcal{C}$ of measures is countable (that is, if $\mathcal{C}$ can be represented as $\mathcal{C} := \{\mu_k : k \in \mathbb{N}\}$), then there exists a predictor which performs well for any $\mu \in \mathcal{C}$. Such a predictor can be obtained as a Bayesian mixture $\rho_S := \sum_{k\in\mathbb{N}} w_k\mu_k$, where w_k are summable positive real weights, and it has very strong predictive properties; in particular, ρ_S predicts every $\mu \in \mathcal{C}$ in total variation distance, as follows from the result of [Blackwell and Dubins(1962)]. Total variation distance measures the difference in (predicted and true) conditional probabilities of all future events, that is, not only the probabilities of the next observations, but also of observations that are arbitrary far off in the future (see formal definitions below). In the context of sequence prediction the measure ρ_S was first studied by [Solomonoff(1978)]. Since then, the idea of taking a convex combination of a finite or countable class of measures (or predictors) to obtain a predictor permeates most of the research on sequential prediction (see, for example, [Cesa-Bianchi and Lugosi(2006)]) and more general learning problems [Hutter(2005), Ryabko and Hutter(2008a)]. In practice it is clear that, on the one hand, countable models are not sufficient, since already the class $\mu_p, p \in [0, 1]$ of Bernoulli i.i.d. processes, where p is the probability of 0, is not countable. On the other hand, prediction in total variation can be too strong to require; predicting probabilities of the next observation may be sufficient, maybe even not on every step but in the Cesaro sense. A key observation here is that a predictor $\rho_S = \sum w_k\mu_k$ may be a good predictor not only when the data is generated by one of the processes μ_k, $k \in \mathbb{N}$, but when it comes from a much larger class. Let us consider

this point in more detail. Fix for simplicity $\mathcal{X} = \{0, 1\}$. The Laplace predictor

$$\lambda(x_{n+1} = 0|x_1, \ldots, x_n) = \frac{\#\{i \leq n : x_i = 0\} + 1}{n + |\mathcal{X}|} \tag{1}$$

predicts any Bernoulli i.i.d. process: although convergence in total variation distance of conditional probabilities does not hold, predicted probabilities of the next outcome converge to the correct ones. Moreover, generalizing the Laplace predictor, a predictor λ_k can be constructed for the class M_k of all k-order Markov measures, for any given k. As was found by [Ryabko(1988)], the combination $\rho_R := \sum w_k \lambda_k$ is a good predictor not only for the set $\cup_{k \in \mathbb{N}} M_k$ of all finite-memory processes, but also for any measure μ coming from a much larger class: that of all stationary measures on $\mathcal{X}^\infty$. Here prediction is possible only in the Cesaro sense (more precisely, ρ_R predicts every stationary process in expected time-average Kullback-Leibler divergence, see definitions below). The Laplace predictor itself can be obtained as a Bayes mixture over all Bernoulli i.i.d. measures with uniform prior on the parameter p (the probability of 0). However, as was observed in [Hutter(2007)] (and as is easy to see), the same (asymptotic) predictive properties are possessed by a Bayes mixture with a countably supported prior which is dense in $[0, 1]$ (e.g. taking $\rho := \sum w_k \delta_k$ where $\delta_k, k \in \mathbb{N}$ ranges over all Bernoulli i.i.d. measures with rational probability of 0). For a given k, the set of k-order Markov processes is parametrized by finitely many $[0, 1]$-valued parameters. Taking a dense subset of the values of these parameters, and a mixture of the corresponding measures, results in a predictor for the class of k-order Markov processes. Mixing over these (for all $k \in \mathbb{N}$) yields, as in [Ryabko(1988)], a predictor for the class of all stationary processes. Thus, for the mentioned classes of processes, a predictor can be obtained as a Bayes mixture of countably many measures in the class. An additional reason why this kind of analysis is interesting is because of the difficulties arising in trying to construct Bayesian predictors for classes of processes that can not be easily parametrized. Indeed, a natural way to obtain a predictor for a class $\mathcal{C}$ of stochastic processes is to take a Bayesian mixture of the class. To do this, one needs to define the structure of a probability space on $\mathcal{C}$. If the class $\mathcal{C}$ is well parametrized, as is the case with the set of all Bernoulli i.i.d. process, then one can integrate with respect to the parametrization. In general, when the problem lacks a natural parametrization, although one can define the structure of the probability space on the set of (all) stochastic process measures in many different ways, the results one can obtain will then be with probability 1 with respect to the prior distribution (see, for example, [Jackson et al.(1999)Jackson, Kalai, and Smorodinsky]). Pointwise consistency cannot be assured (see e.g. [Diaconis and Freedman(1986)]) in this case, meaning that some (well-defined) Bayesian predictors are not consistent on some (large) subset of $\mathcal{C}$. Results with prior probability 1 can be hard to interpret if one is not sure that the structure of the probability space defined on the set $\mathcal{C}$ is indeed a natural one for the problem at hand (whereas if one does have a natural parametrization, then usually results for every value of the parameter can be obtained, as in the case with Bernoulli i.i.d. processes mentioned above). The results of the present work show that when a predictor exists it can indeed be given as a Bayesian predictor, which predicts every (and not almost every) measure in the class, while its support is only a countable set.

A related question is formulated as a question about two individual measures, rather than about a class of measures and a predictor. Namely, one can ask under which conditions one stochastic process predicts another. In [Blackwell and Dubins(1962)] it was shown that if one measure is absolutely continuous with respect to another, than the latter predicts the former (the conditional probabilities converge in a very strong sense). In [Ryabko and Hutter(2007), Ryabko and Hutter(2008b)]

a weaker form of convergence of probabilities (in particular, convergence of expected average KL divergence) is obtained under weaker assumptions.

The results. First, we show that if there is a predictor that performs well for every measure coming from a class $\mathcal{C}$ of processes, then a predictor can also be obtained as a convex combination $\sum_{k\in\mathbb{N}} w_k\mu_k$ for some $\mu_k \in \mathcal{C}$ and some $w_k > 0$, $k \in \mathbb{N}$. This holds if the prediction quality is measured by either total variation distance, or expected average KL divergence: one measure of performance that is very strong, the other rather weak. The analysis for the total variation case relies on the fact that if ρ predicts μ in total variation distance, then μ is absolutely continuous with respect to ρ, so that $\rho(x_{1..n})/\mu(x_{1..n})$ converges to a positive number with μ-probability 1 and with a positive ρ-probability. However, if we settle for a weaker measure of performance, such as expected average KL divergence, measures $\mu \in \mathcal{C}$ are typically singular with respect to a predictor ρ. Nevertheless, since ρ predicts μ we can show that $\rho(x_{1..n})/\mu(x_{1..n})$ decreases subexponentially with n (with high probability or in expectation); then we can use this ratio as an analogue of the density for each time step n, and find a convex combination of countably many measures from $\mathcal{C}$ that has desired predictive properties for each n. Combining these predictors for all n results in a predictor that predicts every $\mu \in \mathcal{C}$ in average KL divergence. The proof techniques developed have a potential to be used in solving other questions concerning sequence prediction, in particular, the general question of how to find a predictor for an arbitrary class $\mathcal{C}$ of measures.

We then exhibit some sufficient conditions on the class $\mathcal{C}$, under which a predictor for all measures in $\mathcal{C}$ exists. It is important to note that none of these conditions relies on a parametrization of any kind. The conditions presented are of two types: conditions on asymptotic behaviour of measures in $\mathcal{C}$, and on their local (restricted to first n observations) behaviour. Conditions of the first type concern separability of $\mathcal{C}$ with respect to the total variation distance and the expected average KL divergence. We show that in the case of total variation separability is a necessary and sufficient condition for the existence of a predictor, whereas in the case of expected average KL divergence it is sufficient but is not necessary.

The conditions of the second kind concern the "capacity" of the sets $\mathcal{C}^n := \{\mu^n : \mu \in \mathcal{C}\}$, $n \in \mathbb{N}$, where μ^n is the measure μ restricted to the first n observations. Intuitively, if $\mathcal{C}^n$ is small (in some sense), then prediction is possible. We measure the capacity of $\mathcal{C}^n$ in two ways. The first way is to find the maximum probability given to each sequence $x_1, \dots, x_n$ by some measure in the class, and then take a sum over $x_1, \dots, x_n$. Denoting the obtained quantity c_n, one can show that it grows polynomially in n for some important classes of processes, such as i.i.d. or Markov processes. We show that, in general, if c_n grows subexponentially then a predictor exists that predicts any measure in $\mathcal{C}$ in expected average KL divergence. On the other hand, exponentially growing c_n are not sufficient for prediction. A more refined way to measure the capacity of $\mathcal{C}^n$ is using a concept of channel capacity from information theory, which was developed for a closely related problem of finding optimal codes for a class of sources. We extend corresponding results from information theory to show that sublinear growth of channel capacity is sufficient for the existence of a predictor, in the sense of expected average divergence. Moreover, the obtained bounds on the divergence are optimal up to an additive logarithmic term.

The rest of the paper is organized as follows. Section 2 introduces the notation and definitions. In Section 3 we show that if any predictor works than there is a Bayesian one that works, while in Section 4 we provide several characterizations of predictable classes of processes. Section 4.1 is concerned with separability, while Section 4.2 analyzes conditions based on local behaviour of measures. Finally, Section 5 provides outlook and discussion.

As running examples that illustrate the results of each section we use countable classes of measures, the family of all Bernoulli i.i.d. processes and that of all stationary processes.

2 Preliminaries

Let $\mathcal{X}$ be a finite set. The notation $x_{1..n}$ is used for $x_1, \dots, x_n$. We consider stochastic processes (probability measures) on $(\mathcal{X}^\infty, \mathcal{F})$ where $\mathcal{F}$ is the sigma-field generated by the cylinder sets $[x_{1..n}]$, $x_i \in \mathcal{X}, n \in \mathbb{N}$, where $[x_{1..n}]$ is the set of all infinite sequences that start with $x_{1..n}$. For a finite set A denote $|A|$ its cardinality. We use $\mathbf{E}_\mu$ for expectation with respect to a measure μ.

Next we introduce the measures of the quality of prediction used in this paper. For two measures μ and ρ we are interested in how different the μ- and ρ-conditional probabilities are, given a data sample $x_{1..n}$. Introduce the *(conditional) total variation* distance

$$v(\mu, \rho, x_{1..n}) := \sup_{A \in \mathcal{F}} |\mu(A|x_{1..n}) - \rho(A|x_{1..n})|.$$

Definition 1. *We say that ρ predicts μ in total variation if*

$$v(\mu, \rho, x_{1..n}) \to 0 \ \mu\text{-a.s.}$$

This convergence is rather strong. In particular, it means that ρ-conditional probabilities of arbitrary far-off events converge to μ-conditional probabilities. Moreover, ρ predicts μ in total variation if [Blackwell and Dubins(1962)] and only if [Kalai and Lehrer(1994)] μ is absolutely continuous with respect to ρ:

Theorem 1 ([Blackwell and Dubins(1962), Kalai and Lehrer(1994)])**.** *If ρ, μ are arbitrary probability measures on $(\mathcal{X}^\infty, \mathcal{F})$, then ρ predicts μ in total variation if and only if μ is absolutely continuous with respect to ρ.*

Thus, for a class $\mathcal{C}$ of measures there is a predictor ρ that predicts every $\mu \in \mathcal{C}$ in total variation if and only if every $\mu \in \mathcal{C}$ has a density with respect to ρ. Although such sets of processes are rather large, they do not include even such basic examples as the set of all Bernoulli i.i.d. processes. That is, there is no ρ that would predict in total variation every Bernoulli i.i.d. process measure δ_p, $p \in [0, 1]$, where p is the probability of 0. Therefore, perhaps for many (if not most) practical applications this measure of the quality of prediction is too strong, and one is interested in weaker measures of performance.

For two measures μ and ρ introduce the *expected cumulative Kullback-Leibler divergence (KL divergence)* as

$$d_n(\mu, \rho) := \mathbf{E}_\mu \sum_{t=1}^{n} \sum_{a \in \mathcal{X}} \mu(x_t = a|x_{1..t-1}) \log \frac{\mu(x_t = a|x_{1..t-1})}{\rho(x_t = a|x_{1..t-1})}, \tag{2}$$

In words, we take the expected (over data) average (over time) KL divergence between μ- and ρ-conditional (on the past data) probability distributions of the next outcome.

Definition 2. *We say that ρ predicts μ in expected average KL divergence if*

$$\frac{1}{n} d_n(\mu, \rho) \to 0.$$

This measure of performance is much weaker, in the sense that it requires good predictions only one step ahead, and not on every step but only on average; also the convergence is not with probability 1 but in expectation. With prediction quality so measured, predictors exist for relatively large classes of measures; most notably, [Ryabko(1988)] provides a predictor which predicts every stationary process in expected average KL divergence. A simple but useful identity that we will need (in the context of sequence prediction introduced also in [Ryabko(1988)]) is the following

$$d_n(\mu,\rho) = -\sum_{x_{1..n}\in\mathcal{X}^n} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})}, \tag{3}$$

where on the right-hand side we have simply the KL divergence between measures μ and ρ restricted to the first n observations.

Thus, the results of this work will be established with respect to two very different measures of prediction quality, one of which is very strong and the other rather weak. This suggests that the facts established reflect some fundamental properties of the problem of prediction, rather than those pertinent to particular measures of performance. On the other hand, it remains open to extend the results below to different measures of performance.

3 Fully nonparametric Bayes predictors

In this section we show that if there is a predictor that predicts every μ in some class $\mathcal{C}$, then there is a Bayesian mixture of countably many elements from $\mathcal{C}$ that predicts every $\mu \in \mathcal{C}$ too. This is established for the two notions of prediction quality that were introduced: total variation and expected average KL divergence. After the theorems we present some examples of families of measures for which predictors exist.

Theorem 2. *Let $\mathcal{C}$ be a set of probability measures on $(\mathcal{X}^\infty, \mathcal{F})$. If there is a measure ρ such that ρ predicts every $\mu \in \mathcal{C}$ in total variation, then there is a sequence $\mu_k \in \mathcal{C}$, $k \in \mathbb{N}$ such that the measure $\nu := \sum_{k\in\mathbb{N}} w_k\mu_k$ predicts every $\mu \in \mathcal{C}$ in total variation, where w_k are any positive weights that sum to 1.*

This relatively simple fact can be proven in different ways, relying on the mentioned equivalence [Blackwell and Dubins(1962), Kalai and Lehrer(1994)] of the statements "ρ predicts μ in total variation distance" and "μ is absolutely continuous with respect to ρ." The proof presented below is not the shortest possible, but it uses ideas and techniques that are then generalized to the case of prediction in expected average KL-divergence, which is more involved, since in all interesting cases all measures $\mu \in \mathcal{C}$ are singular with respect to any predictor that predicts all of them. Another proof of Theorem 2 can be obtained from Theorem 4 in the next section. Yet another way would be to derive it from algebraic properties of the relation of absolute continuity, given in [Plesner and Rokhlin(1946)].

Proof. We break the (relatively easy) proof of this theorem into 3 steps, which will make the proof of the next theorem more understandable.

Step 1: densities. For any $\mu \in \mathcal{C}$, since ρ predicts μ in total variation, by Theorem 1, μ has a density (Radon-Nikodym derivative) f_μ with respect to ρ. Thus, for the set T_μ of all sequences $x_1, x_2, ... \in \mathcal{X}^\infty$ on which $f_\mu(x_{1,2,...}) > 0$ (the limit $\lim_{n\to\infty} \frac{\rho(x_{1..n})}{\mu(x_{1..n})}$ exists and is finite and positive)

we have $\mu(T_\mu) = 1$ and $\rho(T_\mu) > 0$. Next we will construct a sequence of measures $\mu_k \in \mathcal{C}$, $k \in \mathbb{N}$ such that the union of the sets T_{μ_k} has probability 1 with respect to every $\mu \in \mathcal{C}$, and will show that this is a sequence of measures whose existence is asserted in the theorem statement.

Step 2: a countable cover and the resulting predictor. Let $\varepsilon_k := 2^{-k}$ and let $m_1 := \sup_{\mu\in\mathcal{C}} \rho(T_\mu)$. Clearly, $m_1 > 0$. Find any $\mu_1 \in \mathcal{C}$ such that $\rho(T_{\mu_1}) \geq m_1 - \varepsilon_1$, and let $T_1 = T_{\mu_1}$. For $k > 1$ define $m_k := \sup_{\mu\in\mathcal{C}} \rho(T_\mu \backslash T_{k-1})$. If $m_k = 0$ then define $T_k := T_{k-1}$, otherwise find any μ_k such that $\rho(T_{\mu_k}\backslash T_{k-1}) \geq m_k - \varepsilon_k$, and let $T_k := T_{k-1} \cup T_{\mu_k}$. Define the predictor ν as $\nu := \sum_{k\in\mathbb{N}} w_k \mu_k$.

Step 3: ν predicts every $\mu \in \mathcal{C}$. Since the sets T_1, $T_2\backslash T_1, \ldots, T_k\backslash T_{k-1}, \ldots$ are disjoint, we must have $\rho(T_k\backslash T_{k-1}) \to 0$, so that $m_k \to 0$ (since $m_k \leq \rho(T_k\backslash T_{k-1}) + \varepsilon_k \to 0$). Let

$$T := \cup_{k\in\mathbb{N}} T_k.$$

Fix any $\mu \in \mathcal{C}$. Suppose that $\mu(T_\mu\backslash T) > 0$. Since μ is absolutely continuous with respect to ρ, we must have $\delta := \rho(T_\mu\backslash T) > 0$. Then for every $k > 1$ we have

$$m_k = \sup_{\mu'\in\mathcal{C}} \rho(T_{\mu'}\backslash T_{k-1}) \geq \rho(T_\mu\backslash T_{k-1}) \geq \rho(T_\mu\backslash T) = \delta > 0,$$

which contradicts $m_k \to 0$. Thus, we have shown that

$$\mu(T \cap T_\mu) = 1. \tag{4}$$

Let us show that every $\mu \in \mathcal{C}$ is absolutely continuous with respect to ν. Indeed, fix any $\mu \in \mathcal{C}$ and suppose $\mu(A) > 0$ for some $A \in \mathcal{F}$. Then from (4) we have $\mu(A \cap T) > 0$, and, by absolute continuity of μ with respect to ρ, also $\rho(A \cap T) > 0$. Since $T = \cup_{k\in\mathbb{N}} T_k$ we must have $\rho(A \cap T_k) > 0$ for some $k \in \mathbb{N}$. Since on the set T_k the measure μ_k has non-zero density f_{μ_k} with respect to ρ, we must have $\mu_k(A \cap T_k) > 0$. (Indeed, $\mu_k(A \cap T_k) = \int_{A\cap T_k} f_{\mu_k} d\rho > 0$.) Hence,

$$\nu(A \cap T_k) \geq w_k \mu_k(A \cap T_k) > 0,$$

so that $\nu(A) > 0$. Thus, μ is absolutely continuous with respect to ν, and so, by Theorem 1, ν predicts μ in total variation distance. □

Thus, examples of families $\mathcal{C}$ for which there is a ρ that predicts every $\mu \in \mathcal{C}$ in total variation, are limited to families of measures which have a density with respect to some measure ρ. On the one hand, from statistical point of view, such families are rather large: the assumption that the probabilistic law in question has a density with respect to some (nice) measure is a standard one in statistics. It should also be mentioned that such families can easily be uncountable. On the other hand, even such basic examples as the set of all Bernoulli i.i.d. measures does not allow for a predictor that predicts every measure in total variation. Indeed, all these processes are singular with respect to one another; in particular, each of the non-overlapping sets T_p of all sequences which have limiting fraction p of 0s has probability 1 with respect to one of the measures and 0 with respect to all others; since there are uncountably many of these measures, there is no measure ρ with respect to which they all would have a density (since such a measure should have $\rho(T_p) > 0$ for all p) . As it was mentioned, predicting in total variation distance means predicting with arbitrarily growing horizon [Kalai and Lehrer(1994)], while prediction in expected average KL divergence is only concerned with the probabilities of the next observation, and only on time and data average. For the latter measure of prediction quality, consistent predictors exist not only for the class of all Bernoulli processes, but also for the class of all stationary processes [Ryabko(1988)]. The next theorem establishes the result similar to Theorem 2 for expected average KL divergence.

Theorem 3. *Let $\mathcal{C}$ be a set of probability measures on $(\mathcal{X}^\infty, \mathcal{F})$. If there is a measure ρ such that ρ predicts every $\mu \in \mathcal{C}$ in expected average KL divergence, then there exist a sequence $\mu_k \in \mathcal{C}$, $k \in \mathbb{N}$ and a sequence $w_k > 0, k \in \mathbb{N}$, such that $\sum_{k,\in\mathbb{N}} w_k = 1$, and the measure $\nu := \sum_{k\in\mathbb{N}} w_k\mu_k$ predicts every $\mu \in \mathcal{C}$ in expected average KL divergence.*

A difference worth noting with respect to the formulation of Theorem 2 (apart from a different measure of divergence) is in that in the latter the weights w_k can be chosen arbitrarily, while in Theorem 3 this is not the case. In general, the statement "$\sum_{k\in\mathbb{N}} w_k\nu_k$ predicts μ in expected average KL divergence for some choice of w_k, $k \in \mathbb{N}$" does not imply "$\sum_{k\in\mathbb{N}} w'_k\nu_k$ predicts μ in expected average KL divergence for every summable sequence of positive $w'_k, k \in \mathbb{N}$," while the implication trivially holds true if the expected average KL divergence is replaced by the total variation. This is illustrated in the last example of this section. An interesting related question (which is beyond the scope of this paper) is how to chose the weights to optimize the behaviour of a predictor before asymptotic.

The idea of the proof of Theorem 3 is as follows. For every μ and every n we consider the sets T^n_μ of those $x_{1..n}$ on which μ is greater than ρ. These sets have to have (from some n on) a high probability with respect to μ. Then since ρ predicts μ in expected average KL divergence, the ρ-probability of these sets cannot decrease exponentially fast (that is, it has to be quite large). (The sequences $\mu(x_{1..n})/\rho(x_{1..n})$, $n \in \mathbb{N}$ will play the role of densities of the proof of Theorem 2, and the sets T^n_μ the role of sets T_μ on which the density is non-zero.) We then use, for each given n, the same scheme to cover the set $\mathcal{X}^n$ with countably many T^n_μ, as was used in the proof of Theorem 2 to construct a countable covering of the set $\mathcal{X}^\infty$, obtaining for each n a predictor ν_n. Then the predictor ν is obtained as $\sum_{n\in\mathbb{N}} w_n\nu_n$, where the weights decrease subexponentially. The latter fact ensures that, although the weights depend on n, they still play no role asymptotically. The technically most involved part of the proof is to show that the sets T^n_μ in asymptotic have sufficiently large weights in those countable covers that we construct for each n. This is used to demonstrate the implication "if a set has a high μ probability, then its ρ-probability does not decrease too fast, provided some regularity conditions." The proof is broken into the same steps as the (simpler) proof of Theorem 2, to make the analogy explicit and the proof more understandable.

Proof. Define the weights $w_k := wk^{-2}$, where w is the normalizer $6/\pi^2$.
Step 1: densities. Define the sets

$$T^n_\mu := \left\{ x_{1..n} \in \mathcal{X}^n : \mu(x_{1..n}) \geq \frac{1}{n}\rho(x_{1..n}) \right\}. \tag{5}$$

Using Markov's inequality, we derive

$$\mu(\mathcal{X}^n \backslash T^n_\mu) = \mu\left(\frac{\rho(x_{1..n})}{\mu(x_{1..n})} > n\right) \leq \frac{1}{n} E_\mu \frac{\rho(x_{1..n})}{\mu(x_{1..n})} = \frac{1}{n}, \tag{6}$$

so that $\mu(T^n_\mu) \to 1$. (Note that if μ is singular with respect to ρ, as is typically the case, then $\frac{\rho(x_{1..n})}{\mu(x_{1..n})}$ converges to 0 μ-a.e. and one can replace $\frac{1}{n}$ in (5) by 1, while still having $\mu(T^n_\mu) \to 1$.)

Step 2n: a countable cover, time n. Fix an $n \in \mathbb{N}$. Define $m^n_1 := \max_{\mu\in\mathcal{C}} \rho(T^n_\mu)$ (since $\mathcal{X}^n$ are finite all suprema are reached). Find any μ^n_1 such that $\rho^n_1(T^n_{\mu^n_1}) = m^n_1$ and let $T^n_1 := T^n_{\mu^n_1}$. For $k > 1$, let $m^n_k := \max_{\mu\in\mathcal{C}} \rho(T^n_\mu \backslash T^n_{k-1})$. If $m^n_k > 0$, let μ^n_k be any $\mu \in \mathcal{C}$ such that $\rho(T^n_{\mu^n_k} \backslash T^n_{k-1}) = m^n_k$, and let $T^n_k := T^n_{k-1} \cup T^n_{\mu^n_k}$; otherwise let $T^n_k := T^n_{k-1}$. Observe that (for each n) there is only a finite

number of positive m_k^n, since the set $\mathcal{X}^n$ is finite; let K_n be the largest index k such that $m_k^n > 0$. Let

$$\nu_n := \sum_{k=1}^{K_n} w_k \mu_k^n. \tag{7}$$

As a result of this construction, for every $n \in \mathbb{N}$ every $k \le K_n$ and every $x_{1..n} \in T_k^n$ using (5) we obtain

$$\nu_n(x_{1..n}) \ge w_k \frac{1}{n} \rho(x_{1..n}). \tag{8}$$

Step 2: the resulting predictor. Finally, define

$$\nu := \frac{1}{2}\gamma + \frac{1}{2}\sum_{n\in\mathbb{N}} w_n \nu_n, \tag{9}$$

where γ is the i.i.d. measure with equal probabilities of all $x \in \mathcal{X}$ (that is, $\gamma(x_{1..n}) = |\mathcal{X}|^{-n}$ for every $n \in \mathbb{N}$ and every $x_{1..n} \in \mathcal{X}^n$). We will show that ν predicts every $\mu \in \mathcal{C}$, and then in the end of the proof (Step r) we will show how to replace γ by a combination of a countable set of elements of $\mathcal{C}$ (in fact, γ is just a regularizer which ensures that ν-probability of any word is never too close to 0).

Step 3: ν predicts every $\mu \in \mathcal{C}$. Fix any $\mu \in \mathcal{C}$. Introduce the parameters $\varepsilon_\mu^n \in (0,1)$, $n \in \mathbb{N}$, to be defined later, and let $j_\mu^n := 1/\varepsilon_\mu^n$. Observe that $\rho(T_k^n \backslash T_{k-1}^n) \ge \rho(T_{k+1}^n \backslash T_k^n)$, for any $k > 1$ and any $n \in \mathbb{N}$, by definition of these sets. Since the sets $T_k^n \backslash T_{k-1}^n$, $k \in \mathbb{N}$ are disjoint, we obtain $\rho(T_k^n \backslash T_{k-1}^n) \le 1/k$. Hence, $\rho(T_\mu^n \backslash T_j^n) \le \varepsilon_\mu^n$ for some $j \le j_\mu^n$, since otherwise $m_j^n = \max_{\mu\in\mathcal{C}} \rho(T_\mu^n \backslash T_{j_\mu^n}^n) > \varepsilon_\mu^n$ so that $\rho(T_{j_\mu^n+1}^n \backslash T_{j_\mu^n}^n) > \varepsilon_\mu^n = 1/j_\mu^n$, which is a contradiction. Thus,

$$\rho(T_\mu^n \backslash T_{j_\mu^n}^n) \le \varepsilon_\mu^n. \tag{10}$$

We can upper-bound $\mu(T_\mu^n \backslash T_{j_\mu^n}^n)$ as follows. First, observe that

$$\begin{aligned} d_n(\mu,\rho) = &- \sum_{x_{1..n}\in T_\mu^n \cap T_{j_\mu^n}^n} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &- \sum_{x_{1..n}\in T_\mu^n \backslash T_{j_\mu^n}^n} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &- \sum_{x_{1..n}\in \mathcal{X}^n \backslash T_\mu^n} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &= I + II + III. \end{aligned} \tag{11}$$

Then, from (5) we get

$$I \ge -\log n. \tag{12}$$

Observe that for every $n \in \mathbb{N}$ and every set $A \subset \mathcal{X}^n$, using Jensen's inequality we can obtain

$$\begin{aligned} -\sum_{x_{1..n}\in A} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} &= -\mu(A) \sum_{x_{1..n}\in A} \frac{1}{\mu(A)} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &\ge -\mu(A) \log \frac{\rho(A)}{\mu(A)} \ge -\mu(A)\log\rho(A) - \frac{1}{2}. \end{aligned} \tag{13}$$

Thus, from (13) and (10) we get

$$II \geq -\mu(T_\mu^n \backslash T_{j_\mu^n}^n) \log \rho(T_\mu^n \backslash T_{j_\mu^n}^n) - 1/2 \geq -\mu(T_\mu^n \backslash T_{j_\mu^n}^n) \log \varepsilon_\mu^n - 1/2. \tag{14}$$

Furthermore,

$$\begin{aligned} III &\geq \sum_{x_{1..n} \in \mathcal{X}^n \backslash T_\mu^n} \mu(x_{1..n}) \log \mu(x_{1..n}) \geq \mu(\mathcal{X}^n \backslash T_\mu^n) \log \frac{\mu(\mathcal{X}^n \backslash T_\mu^n)}{|\mathcal{X}^n \backslash T_\mu^n|} \\ &\geq -\frac{1}{2} - \mu(\mathcal{X}^n \backslash T_\mu^n) n \log |\mathcal{X}| \geq -\frac{1}{2} - \log |\mathcal{X}|, \end{aligned} \tag{15}$$

where in the second inequality we have used the fact that entropy is maximized when all events are equiprobable, in the third one we used $|\mathcal{X}^n \backslash T_\mu^n| \leq |\mathcal{X}|^n$, while the last inequality follows from (6). Combining (11) with the bounds (12), (14) and (15) we obtain

$$d_n(\mu, \rho) \geq -\log n - \mu(T_\mu^n \backslash T_{j_\mu^n}^n) \log \varepsilon_\mu^n - 1 - \log |\mathcal{X}|,$$

so that

$$\mu(T_\mu^n \backslash T_{j_\mu^n}^n) \leq \frac{1}{-\log \varepsilon_\mu^n} \Big(d_n(\mu, \rho) + \log n + 1 + \log |\mathcal{X}| \Big). \tag{16}$$

Since $d_n(\mu, \rho) = o(n)$, we can define the parameters ε_μ^n in such a way that $-\log \varepsilon_\mu^n = o(n)$ while at the same time the bound (16) gives $\mu(T_\mu^n \backslash T_{j_\mu^n}^n) = o(1)$. Fix such a choice of ε_μ^n. Then, using $\mu(T_\mu^n) \to 1$, we can conclude

$$\mu(\mathcal{X}^n \backslash T_{j_\mu^n}^n) \leq \mu(\mathcal{X}^n \backslash T_\mu^n) + \mu(T_\mu^n \backslash T_{j_\mu^n}^n) = o(1). \tag{17}$$

We proceed with the proof of $d_n(\mu, \nu) = o(n)$. For any $x_{1..n} \in T_{j_\mu^n}^n$ we have

$$\nu(x_{1..n}) \geq \frac{1}{2} w_n \nu_n(x_{1..n}) \geq \frac{1}{2} w_n w_{j_\mu^n} \frac{1}{n} \rho(x_{1..n}) = \frac{w_n w}{2n} (\varepsilon_\mu^n)^2 \rho(x_{1..n}), \tag{18}$$

where the first inequality follows from (9), the second from (8), and in the equality we have used $w_{j_\mu^n} = w/(j_\mu^n)^2$ and $j_\mu^n = 1/\varepsilon_n^\mu$. Next we use the decomposition

$$d_n(\mu, \nu) = -\sum_{x_{1..n} \in T_{j_\mu^n}^n} \mu(x_{1..n}) \log \frac{\nu(x_{1..n})}{\mu(x_{1..n})} - \sum_{x_{1..n} \in \mathcal{X}^n \backslash T_{j_\mu^n}^n} \mu(x_{1..n}) \log \frac{\nu(x_{1..n})}{\mu(x_{1..n})} = I + II. \tag{19}$$

From (18) we find

$$\begin{aligned} I &\leq -\log \Big(\frac{w_n w}{2n} (\varepsilon_\mu^n)^2 \Big) - \sum_{x_{1..n} \in T_{j_\mu^n}^n} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &= (1 + 3 \log n - 2 \log \varepsilon_\mu^n - 2 \log w) + \left(d_n(\mu, \rho) + \sum_{x_{1..n} \in \mathcal{X}^n \backslash T_{j_\mu^n}^n} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \right) \\ &\leq o(n) - \sum_{x_{1..n} \in \mathcal{X}^n \backslash T_{j_\mu^n}^n} \mu(x_{1..n}) \log \mu(x_{1..n}) \\ &\leq o(n) + \mu(\mathcal{X}^n \backslash T_{j_\mu^n}^n) n \log |\mathcal{X}| = o(n), \end{aligned} \tag{20}$$

where in the second inequality we have used $-\log \varepsilon^n_\mu = o(n)$ and $d_n(\mu,\rho) = o(n)$, in the last inequality we have again used the fact that the entropy is maximized when all events are equiprobable, while the last equality follows from (17). Moreover, from (9) we find

$$II \le \log 2 - \sum_{x_{1..n}\in\mathcal{X}^n\setminus T^n_{j^n_\mu}} \mu(x_{1..n}) \log \frac{\gamma(x_{1..n})}{\mu(x_{1..n})} \le 1 + n\mu(\mathcal{X}^n\setminus T^n_{j^n_\mu}) \log|\mathcal{X}| = o(n), \tag{21}$$

where in the last inequality we have used $\gamma(x_{1..n}) = |\mathcal{X}|^{-n}$ and $\mu(x_{1..n}) \le 1$, and the last equality follows from (17).

From (19), (20) and (21) we conclude $\frac{1}{n} d_n(\nu,\mu) \to 0$.

Step r: the regularizer γ. It remains to show that the i.i.d. regularizer γ in the definition of ν (9), can be replaced by a convex combination of a countably many elements from $\mathcal{C}$. Indeed, for each $n \in \mathbb{N}$, denote

$$A_n := \{x_{1..n} \in \mathcal{X}^n : \exists \mu \in \mathcal{C}\ \mu(x_{1..n}) \ne 0\},$$

and let for each $x_{1..n} \in \mathcal{X}^n$ the measure $\mu_{x_{1..n}}$ be any measure from $\mathcal{C}$ such that $\mu_{x_{1..n}}(x_{1..n}) \ge \frac{1}{2}\sup_{\mu\in\mathcal{C}} \mu(x_{1..n})$. Define

$$\gamma'_n(x'_{1..n}) := \frac{1}{|A_n|} \sum_{x_{1..n}\in A_n} \mu_{x_{1..n}}(x'_{1..n}),$$

for each $x'_{1..n} \in A^n$, $n \in \mathbb{N}$, and let $\gamma' := \sum_{k\in\mathbb{N}} w_k \gamma'_k$. For every $\mu \in \mathcal{C}$ we have

$$\gamma'(x_{1..n}) \ge w_n |A_n|^{-1} \mu_{x_{1..n}}(x_{1..n}) \ge \frac{1}{2} w_n |\mathcal{X}|^{-n} \mu(x_{1..n})$$

for every $n \in \mathbb{N}$ and every $x_{1..n} \in A_n$, which clearly suffices to establish the bound $II = o(n)$ as in (21). □

Example: countable classes of measures. A very simple but rich example of a class $\mathcal{C}$ that satisfies the conditions of both the theorems above, is any countable family $\mathcal{C} = \{\mu_k : k \in \mathbb{N}\}$ of measures. In this case, any mixture predictor $\rho := \sum_{k\in\mathbb{N}} w_k \mu_k$ predicts all $\mu \in \mathcal{C}$ both in total variation and in expected average KL divergence. A particular instance that has gained much attention in the literature is the family of all computable measures. Although countable, this family of processes is rather rich. The problem of predicting all computable measures was introduced in [Solomonoff(1978)] where a mixture predictor was proposed.

Example: Bernoulli i.i.d. processes. Consider the class $\mathcal{C}_B = \{\mu_p : p \in [0,1]\}$ of all Bernoulli i.i.d. processes: $\mu_p(x_k = 0) = p$ independently for all $k \in \mathbb{N}$. Clearly, this family is uncountable. Moreover, each set

$$T_p := \{x \in \mathcal{X}^\infty :\ \text{the limiting fraction of 0s in } x \text{ equals } p\},$$

has probability 1 with respect to μ_p and probability 0 with respect to any $\mu_{p'} : p' \ne p$. Since the sets T_p, $p \in [0,1]$ are non-overlapping, there is no measure ρ for which $\rho(T_p) > 0$ for all $p \in [0,1]$. That is, there is no measure ρ with respect to which all μ_p are absolutely continuous. Therefore, by Theorem 1, a predictor that predicts any $\mu \in \mathcal{C}_B$ in total variation does not exist, demonstrating that this notion of prediction is rather strong. However, we know (e.g. [Krichevsky(1993)]) that the Laplace predictor (1) predicts every Bernoulli i.i.d. process in expected average KL divergence (and not only). Hence, Theorem 2 implies that there is a countable mixture predictor for this

family too. Let us find such a predictor. Let $\mu_q : q \in Q$ be the family of all Bernoulli i.i.d. measures with rational probability of 0, and let $\rho := \sum_{q\in Q} w_q\mu_q$, where w_q are arbitrary positive weights that sum to 1. Let μ_p be any Bernoulli i.i.d. process. Let $h(p,q)$ denote the divergence $p\log(p/q) + (1-p)\log(1-p/1-q)$. For each ε we can find a $q \in Q$ such that $h(p,q) < \varepsilon$. Then

$$\frac{1}{n}d_n(\mu_p,\rho) = \frac{1}{n}\mathbf{E}_{\mu_p}\log\frac{\log\mu_p(x_{1..n})}{\log\rho(x_{1..n})} \le \frac{1}{n}\mathbf{E}_{\mu_p}\log\frac{\log\mu_p(x_{1..n})}{w_q\log\mu_q(x_{1..n})}$$
$$= -\frac{\log w_q}{n} + h(p,q) \le \varepsilon + o(1). \quad (22)$$

Since this holds for each ε we conclude that $\frac{1}{n}d_n(\mu_p,\rho) \to 0$ and ρ predicts every $\mu \in \mathcal{C}_B$ in expected average KL divergence.

Example: stationary processes. In [Ryabko(1988)] a predictor ρ_R was constructed which predicts every stationary process $\rho \in \mathcal{C}_S$ in expected average KL divergence. (This predictor is obtained as a mixture of predictors for k-order Markov sources, for all $k \in \mathbb{N}$.) Therefore, Theorem 3 implies that there is also a countable mixture predictor for this family of processes. Such a predictor can be constructed as follows (the proof in this example is based on the proof in [Ryabko and Astola(2006)], Appendix 1). Observe that the family $\mathcal{C}_k$ of k-order stationary binary-valued Markov processes is parametrized by 2^k $[0,1]$-valued parameters: probability of observing 0 after observing $x_{1..k}$, for each $x_{1..k} \in \mathcal{X}^k$. For each $k \in \mathbb{N}$ let μ^k_q, $q \in Q^{2^k}$ be the (countable) family of all stationary k-order Markov processes with rational values of all the parameters. We will show that any predictor $\nu := \sum_{k\in\mathbb{N}}\sum_{q\in Q^{2^k}} w_k w_q \mu^k_q$, where w_k, $k \in \mathbb{N}$ and $w_q, q \in Q^{2^k}$, $k \in \mathbb{N}$ are any sequences of positive real weights that sum to 1, predicts every stationary $\mu \in \mathcal{C}_S$ in expected average KL divergence. For $\mu \in \mathcal{C}_S$ and $k \in \mathbb{N}$ define the k-order conditional Shannon entropy $h_k(\mu) := \mathbf{E}_\mu \log\mu(x_{k+1}|x_{1..k})$. We have $h_{k+1}(\mu) \ge h_k(\mu)$ for every $k \in \mathbb{N}$ and $\mu \in \mathcal{C}_S$, and the limit

$$h_\infty(\mu) := \lim_{k\to\infty} h_k(\mu) \quad (23)$$

is called the limit Shannon entropy, see e.g. [Gallager(1968)]. Fix some $\mu \in \mathcal{C}_S$. It is easy to see that for every $\varepsilon > 0$ and every $k \in \mathbb{N}$ we can find a k-order stationary Markov measure $\mu^k_{q_\varepsilon}$, $q_\varepsilon \in Q^{2^k}$ with rational values of the parameters, such that

$$\mathbf{E}_\mu \log\frac{\mu(x_{k+1}|x_{1..k})}{\mu^k_{q_\varepsilon}(x_{k+1}|x_{1..k})} < \varepsilon. \quad (24)$$

We have

$$\frac{1}{n}d_n(\mu,\nu) \le -\frac{\log w_k w_{q_\varepsilon}}{n} + \frac{1}{n}d_n(\mu,\mu^k_{q_\varepsilon})$$
$$= O(k/n) + \frac{1}{n}\mathbf{E}_\mu\log\mu(x_{1..n}) - \frac{1}{n}\mathbf{E}_\mu\log\mu^k_{q_\varepsilon}(x_{1..n})$$
$$= o(1) + h_\infty(\mu) - \frac{1}{n}\mathbf{E}_\mu\sum_{k=1}^{n}\log\mu^k_{q_\varepsilon}(x_t|x_{1..t-1})$$
$$= o(1) + h_\infty(\mu) - \frac{1}{n}\mathbf{E}_\mu\sum_{t=1}^{k}\log\mu^k_{q_\varepsilon}(x_t|x_{1..t-1}) - \frac{n-k}{n}\mathbf{E}_\mu\log\mu^k_{q_\varepsilon}(x_{k+1}|x_{1..k})$$
$$\le o(1) + h_\infty(\mu) - \frac{n-k}{n}(h_k(\mu) - \varepsilon), \quad (25)$$

where the first inequality is derived analogously to (22), the first equality follows from (3), the second equality follows from the Shannon-McMillan-Breiman theorem (e.g. [Gallager(1968)]), that states that $\frac{1}{n}\log\mu(x_{1..n}) \to h_\infty(\mu)$ in expectation (and a.s.) for every $\mu \in \mathcal{C}_S$, and (3); in the third equality we have used the fact that $\mu^k_{q_\varepsilon}$ is k-order Markov and μ is stationary, whereas the last inequality follows from (24). Finally, since the choice of k and ε was arbitrary, from (25) and (23) we obtain $\lim_{n\to\infty}\frac{1}{n}d_n(\mu,\nu) = 0$.

Example: weights may matter. Finally, we provide an example that illustrates the difference between the formulations of Theorems 2 and 3: in the latter the weights are not arbitrary. We will construct a sequence of measures $\nu_k, k \in \mathbb{N}$, a measure μ, and two sequences of positive weights w_k and w'_k with $\sum_{k\in\mathbb{N}} w_k = \sum_{k\in\mathbb{N}} w'_k = 1$, for which $\nu := \sum_{k\in\mathbb{N}} w_k\nu_k$ predicts μ in expected average KL divergence, but $\nu' := \sum_{k\in\mathbb{N}} w'_k\nu_k$ does not. Let ν_k be a deterministic measure that first outputs k 0s and then only 1s, $k \in \mathbb{N}$. Let $w_k = w/k^2$ with $w = 6/\pi^2$ and $w'_k = 2^{-k}$. Finally, let μ be a deterministic measure that outputs only 0s. We have $d_n(\mu,\nu) = -\log(\sum_{k\geq n} w_k) = O(\log n)$, but $d_n(\mu,\nu') = -\log(\sum_{k\geq n} w'_k) = -\log(2^{-n+1}) = n - 1 \neq o(n)$, proving the claim.

4 Characterizing predictable classes

Knowing that a mixture of a countable subset gives a predictor if there is one, a notion that naturally comes to mind when trying to characterize families of processes for which a predictor exists, is separability. Can we say that there is a predictor for a class $\mathcal{C}$ of measures if and only if $\mathcal{C}$ is separable? Of course, to talk about separability we need a suitable topology on the space of all measures, or at least on $\mathcal{C}$. If the formulated questions were to have a positive answer, we would need a different topology for each of the notions of predictive quality that we consider. Sometimes these measures of predictive quality indeed define a nice enough structure of a probability space, but sometimes they do not. The question whether there exists a topology on $\mathcal{C}$, separability with respect to which is equivalent to the existence of a predictor, is already more vague and less appealing. Nonetheless, in the case of total variation distance we obviously have a candidate topology: that of total variation distance, and indeed separability with respect to this topology is equivalent to the existence of a predictor, as the next theorem shows. This theorem also implies Theorem 2, thereby providing an alternative proof for the latter. In the case of expected average KL divergence the situation is different. While one can introduce a topology based on it, separability with respect to this topology turns out to be a sufficient but not a necessary condition for the existence of a predictor, as is shown in Theorem 5.

4.1 Separability

Definition 3 (unconditional total variation distance)**.** *Introduce the (unconditional) total variation distance*

$$v(\mu,\rho) := \sup_{A\in\mathcal{F}} |\mu(A) - \rho(A)|.$$

Theorem 4. *Let $\mathcal{C}$ be a set of probability measures on $(\mathcal{X}^\infty, \mathcal{F})$. There is a measure ρ such that ρ predicts every $\mu \in \mathcal{C}$ in total variation if and only if $\mathcal{C}$ is separable with respect to the topology of total variation distance. In this case any measure ν of the form $\nu = \sum_{k=1}^\infty w_k\mu_k$, where $\{\mu_k : k \in \mathbb{N}\}$ is any dense countable subset of $\mathcal{C}$ and w_k are any positive weights that sum to 1, predicts every $\mu \in \mathcal{C}$ in total variation.*

Proof. Sufficiency and the mixture predictor. Let $\mathcal{C}$ be separable in total variation distance, and let $\mathcal{D} = \{\nu_k : k \in \mathbb{N}\}$ be its dense countable subset. We have to show that $\nu := \sum_{k\in\mathbb{N}} w_k\nu_k$, where w_k are any positive real weights that sum to 1, predicts every $\mu \in \mathcal{C}$ in total variation. To do this, it is enough to show that $\mu(A) > 0$ implies $\nu(A) > 0$ for every $A \in \mathcal{F}$ and every $\mu \in \mathcal{C}$. Indeed, let A be such that $\mu(A) = \varepsilon > 0$. Since $\mathcal{D}$ is dense in $\mathcal{C}$, there is a $k \in \mathbb{N}$ such that $v(\mu, \nu_k) < \varepsilon/2$. Hence $\nu_k(A) \geq \mu(A) - v(\mu, \nu_k) \geq \varepsilon/2$ and $\nu(A) \geq w_k\nu_k(A) \geq w_k\varepsilon/2 > 0$.

Necessity. For any $\mu \in \mathcal{C}$, since ρ predicts μ in total variation, μ has a density (Radon-Nikodym derivative) f_μ with respect to ρ. We can define L_1 distance with respect to ρ as follows $L_1^\rho(\mu, \nu) = \int_{\mathcal{X}^\infty} |f_\mu - f_\nu| d\rho$. The set of all measures that have a density with respect to ρ is separable with respect to this distance (for example a dense countable subset can be constructed based on measures whose densities are step-functions with finitely many steps, that take only rational values, see e.g. [Kolmogorov and Fomin(1975)]); therefore, its subset $\mathcal{C}$ is also separable. Let $\mathcal{D}$ be any dense countable subset of $\mathcal{C}$. Thus, for every $\mu \in \mathcal{C}$ and every ε there is a $\mu' \in \mathcal{D}$ such that $L_1^\rho(\mu, \mu') < \varepsilon$. For every measurable set A we have

$$|\mu(A) - \mu'(A)| = \left| \int_A f_\mu d\rho - \int_A f_{\mu'} d\rho \right| \leq \int_A |f_\mu - f_{\mu'}| d\rho \leq \int_{\mathcal{X}^\infty} |f_\mu - f_{\mu'}| d\rho < \varepsilon.$$

Therefore, $v(\mu, \mu') = \sup_{A\in\mathcal{F}} |\mu(A) - \mu'(A)| < \varepsilon$ and the set $\mathcal{C}$ is separable in total variation distance. □

Definition 4 (asymptotic KL "distance" D)**.** *Define asymptotic expected average KL divergence between measures μ and ρ as*

$$D(\mu, \rho) = \limsup_{n\to\infty} \frac{1}{n} d_n(\mu, \rho). \tag{26}$$

Theorem 5. *For a set $\mathcal{C}$ of measures, separability with respect to the asymptotic expected average KL divergence D is a sufficient but not a necessary condition for the existence of a predictor:*

(i) If there exists a countable set $\mathcal{D} := \{\nu_k : k \in \mathbb{N}\} \subset \mathcal{C}$ such that for every $\mu \in \mathcal{C}$ and every $\varepsilon > 0$ there is a measure $\mu' \in \mathcal{D}$ such that $D(\mu, \mu') < \varepsilon$, then every measure ν of the form $\nu = \sum_{k=1}^\infty w_k\mu_k$, where w_k are any positive weights that sum to 1, predicts every $\mu \in \mathcal{C}$ in expected average KL divergence.

(ii) There is an uncountable set $\mathcal{C}$ of measures and a measure ν such that ν predicts every $\mu \in \mathcal{C}$ in expected average KL divergence, but $\mu_1 \neq \mu_2$ implies $D(\mu_1, \mu_2) = \infty$ for every $\mu_1, \mu_2 \in \mathcal{C}$; in particular, $\mathcal{C}$ is not separable with respect to D.

Proof. (i) Fix $\mu \in \mathcal{C}$. For every $\varepsilon > 0$ pick $k \in \mathbb{N}$ such that $D(\mu, \nu_k) < \varepsilon$. We have

$$d_n(\mu, \nu) = \mathbf{E}_\mu \log \frac{\mu(x_{1..n})}{\nu(x_{1..n})} \leq \mathbf{E}_\mu \log \frac{\mu(x_{1..n})}{w_k\nu_k(x_{1..n})} = -\log w_k + d_n(\mu, \nu_k) \leq n\varepsilon + o(n).$$

Since this holds for every ε, we conclude $\frac{1}{n} d_n(\mu, \nu) \to 0$.

(ii) Let $\mathcal{C}$ be the set of all deterministic sequences (measures concentrated on just one sequence) such that the number of 0s in the first n symbols is less than $\sqrt{n}$. Clearly, this set is uncountable. It is easy to check that $\mu_1 \neq \mu_2$ implies $D(\mu_1, \mu_2) = \infty$ for every $\mu_1, \mu_2 \in \mathcal{C}$, but the predictor ν given by $\nu(x_n = 0) = 1/n$ independently for different n, predicts every $\mu \in \mathcal{C}$ in expected average KL divergence. □

Examples. Basically, the examples of the preceding section carry over here. Indeed, the example of countable families is trivially also an example of separable (with respect to either of the considered topologies) family. For Bernoulli i.i.d. and k-order Markov processes, the (countable) sets of processes that have rational values of the parameters, considered in the previous section, are dense both in the topology of the parametrization and with respect to the asymptotic average divergence D. It is also easy to check from the arguments presented in the corresponding example of Section 3 that the family of all k-order stationary Markov processes with rational values of the parameters, where we take all $k \in \mathbb{N}$, is dense with respect to D in the set $\mathcal{C}_S$ of all stationary processes, so that $\mathcal{C}_S$ is separable with respect to D. Thus, the sufficient but not necessary condition of separability is satisfied in this case. On the other hand, neither of these latter families is separable with respect to the topology of total variation distance.

4.2 Conditions based on the local behaviour of measures.

Next we provide some sufficient conditions for the existence of a predictor based on local characteristics of the class of measures, that is, measures truncated to the first n observations. First of all, it must be noted that necessary and sufficient conditions cannot be obtained this way. The basic example is that of a family $\mathcal{C}_0$ of all deterministic sequences that are 0 from some time on. This is a countable class of measures which is very easy to predict. Yet the class of measures on $\mathcal{X}^n$ obtained by truncating all measures in $\mathcal{C}_0$ to the first n observation coincides with what would be obtained by truncating all deterministic measures to the first n observation, the latter class being obviously not predictable at all (see also examples below). Nevertheless, considering this kind of local behaviour of measures, one can obtain not only sufficient conditions for the existence of a predictor, but also rates of convergence of the prediction error. It also gives some ideas of how to construct predictors, for the cases when the sufficient conditions obtained are met.

For a class $\mathcal{C}$ of stochastic processes and a sequence $x_{1..n} \in \mathcal{X}^n$ introduce the coefficients

$$c_{x_{1..n}}(\mathcal{C}) := \sup_{\mu \in \mathcal{C}} \mu(x_{1..n}). \tag{27}$$

Define also the normalizer

$$c_n(\mathcal{C}) := \sum_{x_{1..n} \in \mathcal{X}^n} c_{x_{1..n}}(\mathcal{C}). \tag{28}$$

Definition 5 (NML estimate)**.** *The normalized maximum likelihood (e.g. [Krichevsky(1993)]) estimator λ is defined as*

$$\lambda_{\mathcal{C}}(x_{1..n}) := \frac{1}{c_n(\mathcal{C})} c_{x_{1..n}}(\mathcal{C}), \tag{29}$$

for each $x_{1..n} \in \mathcal{X}^n$.

The family $\lambda_{\mathcal{C}}(x_{1..n})$ (indexed by n) in general does not immediately define a stochastic process over $\mathcal{X}^\infty$ ($\lambda_{\mathcal{C}}$ are not consistent for different n); thus, in particular, using average KL divergence for measuring prediction quality would not make sense, since

$$d_n(\mu(\cdot|x_{1..n-1}), \lambda_{\mathcal{C}}(\cdot|x_{1..n-1}))$$

can be negative, as the following example shows.

Example: negative d_n for NML estimates. Let the processes μ_i, $i \in \{1, \dots, 4\}$ be defined on the steps $n = 1, 2$ as follows. $\mu_1(00) = \mu_2(01) = \mu_4(11) = 1$, while $\mu_3(01) = \mu_3(00) = 1/2$. We have

$\lambda_{\mathcal{C}}(1) = \lambda_{\mathcal{C}}(0) = 1/2$, while $\lambda_{\mathcal{C}}(00) = \lambda_{\mathcal{C}}(01) = \lambda_{\mathcal{C}}(11) = 1/3$. If we define $\lambda_{\mathcal{C}}(x|y) = \lambda_{\mathcal{C}}(yx)/\lambda_{\mathcal{C}}(y)$ we get $\lambda_{\mathcal{C}}(1|0) = \lambda_{\mathcal{C}}(0|0) = 2/3$. Then $d_2(\mu_3(\cdot|0), \lambda_{\mathcal{C}}(\cdot|0)) = \log 3/4 < 0$.

Yet, by taking an appropriate mixture, it is still possible to construct a predictor (a stochastic process) based on λ, that predicts all the measures in the class.

Definition 6 (predictor ρ_c)**.** *Let $w := 6/\pi^2$ and let $w_k := \frac{1}{wk^2}$. Define a measure μ_k as follows. On the first k steps it is defined as $\lambda_{\mathcal{C}}$, and for $n > k$ it outputs only zeros with probability 1; so, $\mu_k(x_{1..k}) = \lambda_{\mathcal{C}}(x_{1..k})$ and $\mu_k(x_n = 0) = 1$ for $n > k$. Define the measure ρ_c as*

$$\rho_c = \sum_{k=1}^{\infty} w_k \mu_k. \tag{30}$$

Thus, we have taken the normalized maximum likelihood estimates λ_n for each n and continued them arbitrarily (actually, by a deterministic sequence) to obtain a sequence of measures on $(\mathcal{X}^\infty, \mathcal{F})$ that can be summed.

Theorem 6. *For a class $\mathcal{C}$ of stochastic processes the predictor ρ_c defined above satisfies*

$$\frac{1}{n} d_n(\mu, \rho_c) \le \frac{\log c_n(\mathcal{C})}{n} + O\left(\frac{\log n}{n}\right); \tag{31}$$

in particular, if

$$\log c_n(\mathcal{C}) = o(n). \tag{32}$$

then ρ_c predicts every $\mu \in \mathcal{C}$ in expected average KL divergence.

Proof. Indeed,

$$\begin{aligned}\frac{1}{n} d_n(\mu, \rho_c) = \frac{1}{n}\mathbf{E} \log \frac{\mu(x_{1..n})}{\rho_c(x_{1..n})} &\le \frac{1}{n}\mathbf{E} \log \frac{\mu(x_{1..n})}{w_n \mu_n(x_{1..n})} \\ &\le \frac{1}{n} \log \frac{c_n(\mathcal{C})}{w_n} = \frac{1}{n}(\log c_n(\mathcal{C}) + 2\log n + \log w). \end{aligned} \tag{33}$$

□

Example: i.i.d., finite-memory. To illustrate the applicability of the theorem we first consider the class of i.i.d. processes $\mathcal{C}_B$ over the binary alphabet $\mathcal{X} = \{0, 1\}$. It is easy to see that for each $x_1, \dots, x_n$

$$\sup_{\mu \in \mathcal{C}_B} \mu(x_{1..n}) = (k/n)^k (1 - k/n)^{n-k}$$

where $k = \#\{i \le n : x_i = 0\}$ is the number of 0s in $x_1, \dots, x_n$. For the constants $c_n(\mathcal{C})$ we can derive

$$\begin{aligned} c_n(C) = \sum_{x_{1..n} \in \mathcal{X}^n} \sup_{\mu \in \mathcal{C}_B} \mu(x_{1..n}) &= \sum_{x_{1..n} \in \mathcal{X}^n} (k/n)^k (1 - k/n)^{n-k} \\ &= \sum_{k=0}^{n} \binom{n}{k} (k/n)^k (1 - k/n)^{n-k} \le \sum_{k=0}^{n} \sum_{t=0}^{n} \binom{n}{k} (k/n)^t (1 - k/n)^{n-t} = n + 1, \end{aligned}$$

so that $c_n(C) \le n + 1$.

In general, for the class $\mathcal{C}_k$ of **processes with memory** k over a finite space $\mathcal{X}$ we can get polynomial $c_n(\mathcal{C})$ (see e.g. [Krichevsky(1993)], and also [Ryabko and Hutter(2007)]). Thus, with respect to finite-memory processes, the conditions of Theorem 6 leave ample space for the growth of $c_n(\mathcal{C})$, since (32) allows subexponential growth of $c_n(\mathcal{C})$. Moreover, these conditions are tight, as the following example shows.

Example: exponential coefficients are not sufficient. Observe that the condition (32) cannot be relaxed further, in the sense that exponential coefficients c_n are not sufficient for prediction. Indeed, for the class of all deterministic processes (that is, each process from the class produces some fixed sequence of observations with probability 1) we have $c_n = 2^n$, while obviously for this class a predictor does not exist.

Example: stationary processes. For the set of all stationary processes we can obtain $c_n(C) \geq 2^n/n$ (as is easy to see by considering periodic n-order Markov processes, for each $n \in \mathbb{N}$), so that the conditions of Theorem 6 are not satisfied. This cannot be fixed, since uniform rates of convergence cannot be obtained for this family of processes, as was shown in [Ryabko(1988)].

Optimal rates of convergence. A natural question that arises with respect to the bound (31) is whether it can be matched by a lower bound. This question is closely related to the optimality of the normalized maximum likelihood estimates used in the construction of the predictor. In general, since NML estimates are not optimal, neither are the rates of convergence in (31). To obtain (close to) optimal rates one has to consider a different measure of capacity.

To do so, we make the following connection to a problem in information theory. Let $\mathcal{P}(\mathcal{X}^\infty)$ be the set of all stochastic processes (probability measures) on the space $(\mathcal{X}^\infty, \mathcal{F})$, and let $\mathcal{P}(\mathcal{X})$ be the set of probability distributions over a (finite) set $\mathcal{X}$. For a class $\mathcal{C}$ of measures we are interested in a predictor that has a small (or minimal) worst-case (with respect to the class $\mathcal{C}$) probability of error. Thus, we are interested in the quantity

$$\inf_{\rho \in \mathcal{P}(\mathcal{X}^\infty)} \sup_{\mu \in \mathcal{C}} D(\mu, \rho), \tag{34}$$

where the infimum is taken over all stochastic processes ρ, and D is the asymptotic expected average KL divergence (26). (In particular, we are interested in the conditions under which the quantity (34) equals zero.) This problem has been studied for the case when the probability measures are over a finite set $\mathcal{X}$, and D is replaced simply by the KL divergence d between the measures. Thus, the problem was to find the probability measure ρ (if it exists) on which the following minimax is attained

$$R(A) := \inf_{\rho \in \mathcal{P}(\mathcal{X})} \sup_{\mu \in A} d(\mu, \rho), \tag{35}$$

where $A \subset \mathcal{P}(\mathcal{X})$. This problem is closely related to the problem of finding the best code for the class of sources A, which was its original motivation. The normalized maximum likelihood distribution considered above does not in general lead to the optimum solution for this problem. The optimum solution is obtained through the result that relates the minimax (35) to the so-called channel capacity.

Definition 7 (Channel capacity)**.** *For a set A of measures on a finite set $\mathcal{X}$ the* channel capacity *of A is defined as*

$$C(A) := \sup_{P \in \mathcal{P}_0(A)} \sum_{\mu \in S(P)} P(\mu) d(\mu, \rho_P), \tag{36}$$

where $\mathcal{P}_0(A)$ is the set of all probability distributions on A that have a finite support, $S(P)$ is the (finite) support of a distribution $P \in \mathcal{P}_0(A)$, and $\rho_P = \sum_{\mu \in S(P)} P(\mu)\mu$.

It is shown in [Ryabko(1979), Gallager(1976 (revised 1979))] that $C(A) = R(A)$, thus reducing the problem of finding a minimax to an optimization problem. For probability measures over infinite spaces this result ($R(A) = C(A)$) was generalized by [Haussler(1997)], but the divergence between probability distributions is measured by KL divergence (and not asymptotic average KL divergence), which gives infinite $R(A)$ e.g. already for the class of i.i.d. processes.

However, truncating measures in a class $\mathcal{C}$ to the first n observations, we can use the results about channel capacity to analyze the predictive properties of the class. Moreover, the rates of convergence that can be obtained along these lines are close to optimal. In order to pass from measures minimizing the divergence for each individual n to a process that minimizes the divergence for all n we use the same idea as when constructing the process ρ_c.

Theorem 7. *Let $\mathcal{C}$ be a set of measures on $(\mathcal{X}^\infty, \mathcal{F})$, and let $\mathcal{C}^n$ be the class of measures from $\mathcal{C}$ restricted to $\mathcal{X}^n$. There exists a measure ρ_C such that*

$$\frac{1}{n} d_n(\mu, \rho_C) \le \frac{C(\mathcal{C}^n)}{n} + O\left(\frac{\log n}{n}\right); \tag{37}$$

in particular, if $C(\mathcal{C}^n)/n \to 0$ then ρ_C predicts every $\mu \in \mathcal{C}$ in expected average KL divergence. Moreover, for any measure ρ_C and every $\varepsilon > 0$ there exists $\mu \in \mathcal{C}$ such that

$$\frac{1}{n} d_n(\mu, \rho_C) \ge \frac{C(\mathcal{C}^n)}{n} - \varepsilon.$$

Proof. As shown in [Gallager(1976 (revised 1979))], for each n there exists a sequence ν_k^n, $k \in \mathbb{N}$ of measures on $\mathcal{X}^n$ such that

$$\lim_{k\to\infty} \sup_{\mu\in\mathcal{C}^n} d_n(\mu, \nu_k^n) \to C(\mathcal{C}^n).$$

For each $n \in \mathbb{N}$ find an index k_n such that

$$|\sup_{\mu\in\mathcal{C}^n} d_n(\mu, \nu_{k_n}^n) - C(\mathcal{C}^n)| \le 1.$$

Define the measure ρ_n as follows. On the first n symbols it coincides with $\nu_{k_n}^n$ and $\rho_n(x_m = 0) = 1$ for $m > n$. Finally, set $\rho_C = \sum_{n=1}^\infty w_n \rho_n$, where $w_k = \frac{1}{wn^2}, w = 6/\pi^2$. We have to show that $\lim_{n\to\infty} \frac{1}{n} d_n(\mu, \rho_C) = 0$ for every $\mu \in \mathcal{C}$. Indeed, similarly to (33), we have

$$\begin{aligned}\frac{1}{n} d_n(\mu, \rho_C) &= \frac{1}{n} \mathbf{E}_\mu \log \frac{\mu(x_{1..n})}{\rho_C(x_{1..n})} \\ &\le \frac{\log w_k^{-1}}{n} + \frac{1}{n} \mathbf{E}_\mu \log \frac{\mu(x_{1..n})}{\rho_n(x_{1..n})} \le \frac{\log w + 2\log n}{n} + \frac{1}{n} d_n(\mu, \rho_n) \\ &\le o(1) + \frac{C(\mathcal{C}^n)}{n}. \end{aligned} \tag{38}$$

The second statement follows from the fact [Ryabko(1979), Gallager(1976 (revised 1979))] that $C(\mathcal{C}^n) = R(\mathcal{C}^n)$ (cf. (35)). □

Thus, if the channel capacity $C(\mathcal{C}^n)$ grows sublinearly, a predictor can be constructed for the class of processes $\mathcal{C}$. In this case the problem of constructing the predictor is reduced to finding

the channel capacities for different n and finding the corresponding measures on which they are attained or approached.

Examples. For the class of all Bernoulli i.i.d. processes, the channel capacity $C(\mathcal{C}_B^n)$ is known to be $O(\log n)$ [Krichevsky(1993)]. For the family of all stationary processes it is $O(n)$, so that the conditions of Theorem 7 are satisfied for the former but not for the latter.

We also remark that the requirement of a sublinear channel capacity cannot be relaxed, in the sense that a linear channel capacity is not sufficient for prediction, since it is the maximal possible capacity for a set of measures on $\mathcal{X}^n$, achieved e.g. on the set of all measures, or on the set of all deterministic sequences.

5 Discussion

The first possible extension of the results of the paper that comes to mind is to find out whether the same holds for other measures of performance, such as prediction in KL divergence without time-averaging, or with probability 1 rather then in expectation, or with respect to other measures of prediction error, such as absolute distance. (See [Ryabko and Hutter(2007)] for a discussion of different measures of performance and relations between them.) Maybe the same results can be obtained in more general formulations, for example, using f-divergences of [Csiszar(1967)].

More generally, the questions we addressed in this work are a part of a larger problem: given an arbitrary class $\mathcal{C}$ of stochastic processes, find the best predictor for it. We have considered two subproblems: first, in which form to look for a predictor if one exists. Here we have shown that if any predictor works then a Bayesian one works too. The second one is to characterize families of processes for which a predictor exists. Here we have analyzed what the notion of separability furnishes in this respect, as well as identified some simple sufficient conditions based on the local behaviour of measures in the class. Another approach would be to identify the conditions which two measures μ and ρ have to satisfy in order for ρ to predict μ. For prediction in total variation such conditions have been identified [Blackwell and Dubins(1962), Kalai and Lehrer(1994)] and, in particular, in the context of the present work, they turn out to be very useful. [Kalai and Lehrer(1994)] also provide some characterization for the case of a weaker notion of prediction: difference between conditional probabilities of the next (several) outcomes (weak merging of opinions). In [Ryabko and Hutter(2008b)] some sufficient conditions are found for the case of prediction in expected average KL divergence, and prediction in average KL divergence with probability 1. Of course, another very natural approach to the general problem posed above is to try and find predictors (in the form of algorithms) for some particular classes of processes which are of practical interest. Towards this end, we have found a rather simple form that some solution to this question has if a solution exists: a Bayesian predictor whose prior is concentrated on a countable set. We have also identified some sufficient conditions under which a predictor can actually be constructed (e.g. using NML estimates). However, the larger question of how to construct an optimal predictor for an arbitrary given family of processes, remains open.

Taking an even more general perspective, one can consider the problem of finding the best response to the actions of a (stochastic) environment, which itself responds to the actions of a learner. Allowing into consideration environments that change their behaviour in response to the action of the learner, clearly makes the problem much more difficult, but it also dramatically extends the range of applications. For this general problem one can pose the same questions: given a set $\mathcal{C}$ of environments, how can we construct a learner that is (asymptotically) optimal if any

environment from $\mathcal{C}$ is chosen to generate the data? One can consider Bayesian learners for this formulation too [Hutter(2005)]; it would be interesting to find out whether one can show that when there is an learner which is optimal in every environment from $\mathcal{C}$, then there is a Bayesian learner with a countably supported prior that has this property too.

References

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

[Cesa-Bianchi and Lugosi(2006)] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. ISBN 0521841089.

[Csiszar(1967)] I. Csiszar. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar*, 2:299–318, 1967.

[Diaconis and Freedman(1986)] P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *Annals of Statistics*, 14(1):1–26, 1986.

[Gallager(1968)] R. G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, New York, NY, USA, 1968.

[Gallager(1976 (revised 1979))] R. G. Gallager. Source coding with side information and universal coding. Technical Report LIDS-P-937, M.I.T., 1976 (revised 1979).

[Haussler(1997)] D. Haussler. A general minimax result for relative entropy. *IEEE Trans. on Information Theory*, 43(4):1276–1280, 1997.

[Hutter(2005)] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005.

[Hutter(2007)] M. Hutter. On universal prediction and Bayesian confirmation. *Theoretical Computer Science*, 348(1):33–48, 2007.

[Jackson et al.(1999)Jackson, Kalai, and Smorodinsky] M. Jackson, E. Kalai, and R. Smorodinsky. Bayesian representation of stochastic processes under learning: de Finetti revisited. *Econometrica*, 67(4):875–794, 1999.

[Kalai and Lehrer(1994)] E. Kalai and E. Lehrer. Weak and strong merging of opinions. *Journal of Mathematical Economics*, 23:73–86, 1994.

[Kolmogorov and Fomin(1975)] A. N. Kolmogorov and S. V. Fomin. *Elements of the Theory of Functions and Functional Analysis*. Dover, 1975.

[Krichevsky(1993)] R. Krichevsky. *Universal Compression and Retrival*. Kluwer Academic Publishers, 1993.

[Plesner and Rokhlin(1946)] A.I. Plesner and V.A. Rokhlin. Spectral theory of linear operators, II. *Uspekhi Matematicheskikh Nauk*, 1:71–191, 1946.

[Ryabko(1979)] B. Ryabko. Coding of a source with unknown but ordered probabilities. *Problems of Information Transmission*, 15(2):134–138, 1979.

[Ryabko(1988)] B. Ryabko. Prediction of random sequences and universal coding. *Problems of Information Transmission*, 24:87–96, 1988.

[Ryabko and Astola(2006)] B. Ryabko and J. Astola. Universal codes as a basis for time series testing. *Statistical Methodology*, 3:375–397, 2006.

[Ryabko(2008)] D. Ryabko. Some sufficient conditions on an arbitrary class of stochastic processes for the existence of a predictor. In *Proc. 19th International Conf. on Algorithmic Learning Theory (ALT'08)*, LNAI 5254, pages 169–182, Budapest, Hungary, 2008. Springer, Berlin.

[Ryabko(2009)] D. Ryabko. Characterizing predictable classes of processes. In A. Ng J. Bilmes, editor, *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI'09)*, Montreal, Canada, 2009.

[Ryabko and Hutter(2007)] D. Ryabko and M. Hutter. On sequence prediction for arbitrary measures. In *Proc. 2007 IEEE International Symposium on Information Theory*, pages 2346–2350, Nice, France, 2007. IEEE. ISBN 1-4244-1429-6.

[Ryabko and Hutter(2008a)] D. Ryabko and M. Hutter. On the possibility of learning in reactive environments with arbitrary dependence. *Theoretical Computer Science*, 405(3):274–284, 2008a.

[Ryabko and Hutter(2008b)] D. Ryabko and M. Hutter. Predicting non-stationary processes. *Applied Mathematics Letters*, 21(5):477–482, 2008b.

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