Characterizing predictable classes of processes

Daniil Ryabko

Characterizing predictable classes of processes

Daniil Ryabko
INRIA Lille-Nord Europe,
daniil@ryabko.net

May 26, 2009

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 class $\mathcal{C}$ of stochastic processes. We are interested in predictors ρ whose conditional probabilities converge to the "true" μ-conditional probabilities if any $\mu \in \mathcal{C}$ is chosen to generate the data. We show that if such a predictor exists, then a predictor can also be obtained as a convex combination of a countably many elements of $\mathcal{C}$. In other words, it can be obtained as a Bayesian predictor whose prior is concentrated on a countable set. This result is established for two very different measures of performance of prediction, one of which is very strong, namely, total variation, and the other is very weak, namely, prediction in expected average Kullback-Leibler divergence.

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 $x \in \mathcal{X}$, 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 set 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 in that we we provide a specific form in which to look for a solution to the general problem. More precisely, we show that if a predictor exists, then a predictor can also be obtained as a weighted sum of a 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.

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, bioninformatics, 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. In order to do this, the first step is to understand for which classes of problems (i.e. sets of measures generating the data) a predictor exists.

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

some related topics in AI [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 $\lambda(x_{n+1} = 0|x_1, \dots, x_n) = \frac{\#\{i \le n : x_i = 0\}+1}{n+|\mathcal{X}|}$ 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 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 processes 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]), while pointwise consistency cannot be assured (see e.g. [Diaconis and Freedman, 1986]). Re-

sults 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 countable.

The results. Here 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 hight probability), and 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 then 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.

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 *total variation* distance

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

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.

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

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

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

$$d_n(\mu,\rho) = -\sum_{x_{1..n}\in\mathcal{X}^n} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})}, \tag{2}$$

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 Main results

Theorem 1. *Let $\mathcal{C}$ be a set of probability measures on $\mathcal{X}^\infty$. 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 equivalence of the statements "ρ predicts μ in total variation distance" and "μ is absolutely continuous with respect to ρ." The proof presented below uses techniques that can be then generalized to the case of prediction in expected average KL-divergence, where in all interesting cases all measures $\mu \in \mathcal{C}$ are singular with respect to any predictor that predicts all of them. The idea of the proof of Theorem 1 is as follows. For each measure $\mu \in \mathcal{C}$ we find the set T_μ of sequences $x_1, x_2, \dots$ on which the density of μ with respect to ρ exists and is non-zero. Such a set has μ-probability 1, and, by absolute continuity, a positive ρ-probability. The idea is then to cover the union $\cup_{\mu\in\mathcal{C}} T_\mu$ with countably many of these sets, and then construct a new predictor as a sum of the corresponding measures. To find this countable collection of sets T_μ, we first find a largest (up to an ε_1) one with respect ρ, then the one who has a largest (up to an ε_2) part not covered by the first set, and so on (where ε_k are decreasing). Then we show that any strictly convex combination of the resulting sequence of measures has the property that any measure in $\mathcal{C}$ is absolutely continuous with respect to it.

Proof. We break the (relatively easy) proof of this theorem into 3 steps, which will make the (more involved) proof of the next theorem more understandable. *Step 1: densities.* 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. Thus, for the set T_μ of all sequences $x_1, x_2, ... \in \mathcal{X}^\infty$ on which $f_\mu(x_{1,2,\dots}) > 0$ (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, $T_2\backslash T_1, \dots, T_k\backslash T_{k-1}, \dots$ are disjoint, we must have $\rho(T_k\backslash T_{k-1}) \to 0$, so that $m_k \to 0$. 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

$$m_k = \sup_{\mu'\in\mathcal{C}} \rho(T_{\mu'}\backslash T_{k-1}) \geq \rho(T_\mu\backslash T_{k-1}) \geq \delta > 0,$$

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

$$\mu(T \cap T_\mu) = 1. \tag{3}$$

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 (3) 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 ν predicts μ in total variation distance. □

Theorem 2. *Let $\mathcal{C}$ be a set of probability measures on $\mathcal{X}^\infty$. If there is a measure ρ such that ρ predicts every $\mu \in \mathcal{C}$ in expected average KL divergence, 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 expected average KL divergence, where w_k are some positive weights.*

A difference worth noting with respect to the formulation of Theorem 1 (apart from a different measure of divergence) is in that in the latter the weights w_k can be chosen arbitrarily, while in Theorem 2 they can not. 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. 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 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 1, 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 1 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 1, 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

$$T^n_\mu := \left\{ x_{1..n} \in \mathcal{X}^n : \mu(x_{1..n}) \geq \frac{1}{n}\rho(x_{1..n}) \right\}. \tag{4}$$

Using Markov's inequality, we derive

$$\mu(\mathcal{X}^n \backslash T^n_\mu) = \mu\left(\frac{\rho(x_{1..n})}{\mu(x_{1..n})} > n\right) \leq \frac{1}{n} E_\mu \frac{\rho(x_{1..n})}{\mu(x_{1..n})} = \frac{1}{n}, \tag{5}$$

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 (4) 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^n_k, since the set $\mathcal{X}^n$ is finite; let K_n be the largest index k such that $m^n_k > 0$. Let

$$\nu_n := \sum_{k=1}^{K_n} w_k \mu^n_k. \tag{6}$$

As a result of this construction, for every $n \in \mathbb{N}$ every $k \leq K_n$ and every $x_{1..n} \in T^n_k$ using (4) we obtain

$$\nu_n(x_{1..n}) \geq w_k \frac{1}{n}\rho(x_{1..n}). \tag{7}$$

Step 2: the resulting predictor. Finally, define

$$\nu := \frac{1}{2}\gamma + \frac{1}{2}\sum_{n\in\mathbb{N}} w_n \nu_n, \tag{8}$$

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 $\varepsilon^n_\mu \in (0,1)$, $n \in \mathbb{N}$, to be defined later, and let $j^n_\mu := 1/\varepsilon^n_\mu$. Observe

that $\rho(T_k^n \backslash T_{k-1}^n) \geq \rho(T_{k+1}^n \backslash T_k^n)$, 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 $\rho(T_k^n \backslash T_{k-1}^n) \leq 1/k$. Hence, $\rho(T_\mu^n \backslash T_j^n) \leq \varepsilon_\mu^n$ for some $j \leq j_\mu^n$, 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,

$$\rho(T_\mu^n \backslash T_{j_\mu^n}^n) \leq \varepsilon_\mu^n. \tag{9}$$

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

$$\begin{aligned} d_n(\mu, \rho) = &- \sum_{x_{1..n} \in T_\mu^n \cap T_{j_\mu^n}^n} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &- \sum_{x_{1..n} \in T_\mu^n \backslash T_{j_\mu^n}^n} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &- \sum_{x_{1..n} \in \mathcal{X}^n \backslash T_\mu^n} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &= I + II + III. \end{aligned} \tag{10}$$

Then, from (4) we get

$$I \geq -\log n. \tag{11}$$

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

$$\begin{aligned} -\sum_{x_{1..n} \in A} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} &= -\mu(A) \sum_{x_{1..n} \in A} \frac{1}{\mu(A)} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &\geq -\mu(A) \log \frac{\rho(A)}{\mu(A)} \geq -\mu(A) \log \rho(A) - \frac{1}{2}. \end{aligned} \tag{12}$$

Thus, from (12) and (9) we get

$$II \geq -\mu(T_\mu^n \backslash T_{j_\mu^n}^n) \log \rho(T_\mu^n \backslash T_{j_\mu^n}^n) - 1/2 \geq -\mu(T_\mu^n \backslash T_{j_\mu^n}^n) \log \varepsilon_\mu^n - 1/2. \tag{13}$$

Furthermore,

$$\begin{aligned} III \geq \sum_{x_{1..n} \in \mathcal{X}^n \backslash T_\mu^n} \mu(x_{1..n}) \log \mu(x_{1..n}) &\geq \mu(\mathcal{X}^n \backslash T_\mu^n) \log \frac{\mu(\mathcal{X}^n \backslash T_\mu^n)}{|\mathcal{X}^n \backslash T_\mu^n|} \\ &\geq -\frac{1}{2} - \mu(\mathcal{X}^n \backslash T_\mu^n) n \log |\mathcal{X}| \geq -\frac{1}{2} - \log |\mathcal{X}|, \end{aligned} \tag{14}$$

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 (5). Combining (10) with the bounds (11), (13) and (14) 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

$$\mu(T_\mu^n \backslash T_{j_\mu^n}^n) \leq \frac{1}{-\log \varepsilon_\mu^n} \Big(d_n(\mu,\rho) + \log n + 1 + \log |\mathcal{X}| \Big). \tag{15}$$

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 (15) 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

$$\mu(\mathcal{X}^n \backslash T_{j_\mu^n}^n) \leq \mu(\mathcal{X}^n \backslash T_\mu^n) + \mu(T_\mu^n \backslash T_{j_\mu^n}^n) = o(1). \tag{16}$$

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

$$\nu(x_{1..n}) \geq \frac{1}{2} w_n \nu_n(x_{1..n}) \geq \frac{1}{2} w_n w_{j_\mu^n} \frac{1}{n} \rho(x_{1..n}) = \frac{w_n w}{2n} (\varepsilon_\mu^n)^2 \rho(x_{1..n}), \tag{17}$$

where the first inequality follows from (8), the second from (7), 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

$$\begin{aligned} d_n(\mu,\nu) = &- \sum_{x_{1..n} \in T_{j_\mu^n}^n} \mu(x_{1..n}) \log \frac{\nu(x_{1..n})}{\mu(x_{1..n})} \\ &- \sum_{x_{1..n} \in \mathcal{X}^n \backslash T_{j_\mu^n}^n} \mu(x_{1..n}) \log \frac{\nu(x_{1..n})}{\mu(x_{1..n})} = I + II. \end{aligned} \tag{18}$$

From (17) we find

$$\begin{aligned} I &\leq -\log \Big(\frac{w_n w}{2n} (\varepsilon_\mu^n)^2 \Big) - \sum_{x_{1..n} \in T_{j_\mu^n}^n} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &= (1 + 3\log n - 2\log \varepsilon_\mu^n - 2\log w) + \left(d_n(\mu,\rho) + \sum_{x_{1..n} \in \mathcal{X}^n \backslash T_{j_\mu^n}^n} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \right) \\ &\leq o(n) - \sum_{x_{1..n} \in \mathcal{X}^n \backslash T_{j_\mu^n}^n} \mu(x_{1..n}) \log \mu(x_{1..n}) \\ &\leq o(n) + \mu(\mathcal{X}^n \backslash T_{j_\mu^n}^n) n \log |\mathcal{X}| = o(n), \end{aligned} \tag{19}$$

where in the second inequality we have used $-\log \varepsilon_\mu^n = o(n)$ 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 (16). Moreover, from (8) we find

$$\begin{aligned} II &\leq \log 2 - \sum_{x_{1..n} \in \mathcal{X}^n \backslash T_{j_\mu^n}^n} \mu(x_{1..n}) \log \frac{\gamma(x_{1..n})}{\mu(x_{1..n})} \\ &\leq 1 + n\mu(\mathcal{X}^n \backslash T_{j_\mu^n}^n) \log |\mathcal{X}| = o(n), \end{aligned} \tag{20}$$

where in the last inequality we have used $\gamma(x_{1..n}) = |\mathcal{X}|^{-n}$ and $\mu(x_{1..n}) \leq 1$, and the last equality follows from (16).

From (18), (19) and (20) 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. regularlizer γ in the definition of ν (8), can be replaced by a convex combination of a countably many elements from $\mathcal{C}$. Indeed, for each $n \in \mathbb{N}$, denote

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

and let $\mu_{x_{1..n}} := \operatorname{argmax}_{\mu \in \mathcal{C}} \mu(x_{1..n})$ for each $x_{1..n} \in \mathcal{X}^n$. 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

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

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 (20). □

4 Discussion

For two measures of quality of prediction that we have considered, namely, total variation distance and expected average KL divergence, we have shown that if a prediction for a class $\mathcal{C}$ of measures exists, then a predictor can also be obtained as a Bayesian mixture over a countable subset of $\mathcal{C}$. The first possible extension of these results 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. Maybe the same results can be obtained in more general formulations, such as 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. One can approach this problem from other sides. For example, the first question one may wish to address is for which classes of processes a predictor exists; see [Ryabko, 2008] for some sufficient conditions, such as separability of the class $\mathcal{C}$. Another approach is 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 provides 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, the contribution of this work is in providing a specific form that some solution to this question has to have, if a solution exists: a Bayesian predictor whose prior is concentrated on a countable set. This is perhaps a rather simple form, which may be useful for constructing practical algorithms.

References

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

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

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

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

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

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

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

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

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

[Ryabko, 2008] Ryabko, D. (2008). 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.

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

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

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