Journal of Machine Learning Research 12 (2011) 1-21 Submitted 12/10; Published 6/11

On the Relation between Realizable and Nonrealizable Cases of the Sequence Prediction Problem

Daniil Ryabko DANIIL@RYABKO.NET
INRIA Lille-Nord Europe,
40, avenue Halley,
Parc Scientifique de la Haute Borne
59650 Villeneuve d'Ascq, France

Editor: Nicolò Cesa-Bianchi

Abstract

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, one is required to give conditional probabilities of the next observation. The realizable case is when the measure μ belongs to an arbitrary but known class $\mathcal{C}$ of process measures. The non-realizable case is when μ is completely arbitrary, but the prediction performance is measured with respect to a given set $\mathcal{C}$ of process measures. We are interested in the relations between these problems and between their solutions, as well as in characterizing the cases when a solution exists and finding these solutions. We show that if the quality of prediction is measured using the total variation distance, then these problems coincide, while if it is measured using the expected average KL divergence, then they are different. For some of the formalizations we also show that when a solution exists, it can be obtained as a Bayes mixture over a countable subset of $\mathcal{C}$. We also obtain several characterization of those sets $\mathcal{C}$ for which solutions to the considered problems exist. As an illustration to the general results obtained, we show that a solution to the non-realizable case of the sequence prediction problem exists for the set of all finite-memory processes, but does not exist for the set of all stationary processes. It should be emphasized that the framework is completely general: the processes measures considered are not required to be i.i.d., mixing, stationary, or to belong to any parametric family.

Keywords: Sequence Prediction, Time Series, Online Prediction, Realizable sequence prediction, Non-realizable sequence prediction.

1. Introduction

A sequence $x_1, \dots, x_n, \dots$ of discrete-valued observations (where x_i belong to a finite set $\mathcal{X}$) is generated according to some unknown probabilistic law (measure). That is, μ is a probability measure on the space $\Omega = (\mathcal{X}^\infty, \mathcal{B})$ of one-way infinite sequences (here $\mathcal{B}$ is the usual Borel σ-algebra). After each new outcome x_n is revealed, one is required to predict conditional *probabilities* of the next observation $x_{n+1} = a$, $a \in \mathcal{X}$, given the past $x_1, \dots, x_n$. Since a predictor ρ is required to give conditional probabilities $\rho(x_{n+1} = a | x_1, \dots, x_n)$ for all possible histories $x_1, \dots, x_n$, it defines itself a probability measure on the space Ω of one-way infinite sequences. In other words, a probability measure on Ω can be considered both as a data-generating mechanism and as a predictor.

Therefore, given a set $\mathcal{C}$ of probability measures on Ω, one can ask two kinds of questions about $\mathcal{C}$. First, does there exist a predictor ρ, whose forecast probabilities converge (in a certain sense) to the μ-conditional probabilities, if an arbitrary $\mu \in \mathcal{C}$ is chosen to generate the data? Here we assume that the "true" measure that generates the data belongs to the set $\mathcal{C}$ of interest, and would like to construct a predictor that predicts all measures in $\mathcal{C}$. The second type of questions is as follows: does there exist a predictor that predicts at least as well as any predictor $\rho \in \mathcal{C}$, if the measure that generates the data comes possibly from outside of $\mathcal{C}$? Thus, here we consider elements of $\mathcal{C}$ as predictors, and we would like to combine their predictive properties, if this is possible. Note that in this setting the two questions above concern the same object: a set $\mathcal{C}$ of probability measures on Ω.

Each of these two questions, the realizable and the non-realizable one, have enjoyed much attention in the literature; the setting for the non-realizable case is usually slightly different, which is probably why it has not (to the best of the author's knowledge) been studied as another facet of the realizable case. The realizable case traces back to Laplace, who has considered the problem of predicting outcomes of a series of independent tosses of a biased coin. That is, he has considered the case when the set $\mathcal{C}$ is that of all i.i.d. process measures. Other classical examples studied are the set of all computable (or semi-computable) measures (Solomonoff, 1978), the set of k-order Markov and finite-memory processes (e.g., Krichevsky, 1993) and the set of all stationary processes (Ryabko, 1988). The general question of finding predictors for an arbitrary given set $\mathcal{C}$ of process measures has been addressed in (Ryabko and Hutter, 2007, 2008; Ryabko, 2010a); the latter work shows that when a solution exists it can be obtained as a Bayes mixture over a countable subset of $\mathcal{C}$.

The non-realizable case is usually studied in a slightly different, non-probabilistic, setting. We refer to (Cesa-Bianchi and Lugosi, 2006) for a comprehensive overview. It is usually assumed that the observed sequence of outcomes is an arbitrary (deterministic) sequence; it is required not to give conditional probabilities, but just deterministic guesses (although these guesses can be selected using randomisation). Predictions result in a certain loss, which is required to be small as compared to the loss of a given set of reference predictors (experts) $\mathcal{C}$. The losses of the experts and the predictor are observed after each round. In this approach, it is mostly assumed that the set $\mathcal{C}$ is finite or countable. The main difference with the formulation considered in this work is that we require a predictor to give probabilities, and thus the loss is with respect to something never observed (probabilities, not outcomes). The loss itself is not completely observable in our setting. In this sense our non-realizable version of the problem is more difficult. Assuming that the data generating mechanism is probabilistic, even if it is completely unknown, makes sense in such problems as, for example, game playing, or market analysis. In these cases one may wish to assign smaller loss to those models or experts who give probabilities closer to the correct ones (which are never observed), even though different probability forecasts can often result in the same action. Aiming at predicting probabilities of outcomes also allows us to abstract from the actual use of the predictions (for example, making bets) and thus from considering losses in a general form; instead, we can concentrate on those forms of loss that are more convenient for the analysis. In this latter respect, the problems we consider are easier than those considered in prediction with expert advice. (However, in principle, nothing restricts us to considering the simple losses that we chose; they are just

a convenient choice.) Noteworthy, the probabilistic approach also makes the machinery of probability theory applicable, hopefully making the problem easier. A reviewer suggested the following summary explanation of the difference between the non-realizable problems of this work and prediction with expert advice: the latter is prequential (in the sense of Dawid, 1992), whereas the former is not.

In this work we consider two measures of the quality of prediction. The first one is the total variation distance, which measures the difference between the forecast and the "true" conditional probabilities of all future events (not just the probability of the next outcome). The second one is expected (over the data) average (over time) Kullback-Leibler divergence. Requiring that predicted and true probabilities converge in total variation is very strong; in particular, this is possible if (Blackwell and Dubins, 1962) and only if (Kalai and Lehrer, 1994) the process measure generating the data is absolutely continuous with respect to the predictor. The latter fact makes the sequence prediction problem relatively easy to analyse. Here we investigate what can be paralleled for the other measure of prediction quality (average KL divergence), which is much weaker, and thus allows for solutions for the cases of much larger sets $\mathcal{C}$ of process measures (considered either as predictors or as data generating mechanisms).

Having introduced our measures of prediction quality, we can further break the non-realizable case into two problems. The first one is as follows. Given a set $\mathcal{C}$ of predictors, we want to find a predictor whose prediction error converges to zero if there is at least one predictor in $\mathcal{C}$ whose prediction error converges to zero; we call this problem simply the "non-realizable" case, or Problem 2 (leaving the name "Problem 1" to the realizable case). The second non-realizable problem is the "fully agnostic" problem: it is to make the prediction error asymptotically as small as that of the best (for the given process measure generating the data) predictor in $\mathcal{C}$ (we call this Problem 3). Thus, we now have three problems about a set of process measures $\mathcal{C}$ to address.

We show that if the quality of prediction is measured in total variation, then all the three problems coincide: any solution to any one of them is a solution to the other two. For the case of expected average KL divergence, all the three problems are different: the realizable case is strictly easier than non-realizable (Problem 2), which is, in turn, strictly easier than the fully agnostic case (Problem 3). We then analyse which results concerning prediction in total variation can be transferred to which of the problems concerning prediction in average KL divergence. It was shown in (Ryabko, 2010a) that, for the realizable case, if there is a solution for a given set of process measures $\mathcal{C}$, then a solution can also be obtained as a Bayesian mixture over a countable subset of $\mathcal{C}$; this holds both for prediction in total variation and in expected average KL divergence. Here we show that this result also holds true for the (non-realizable) case of Problem 2, for prediction in expected average KL divergence. We do not have an analogous result for Problem 3 (and, in fact, conjecture that the opposite statement holds true). However, for the fully agnostic case of Problem 3, we show that separability with respect to a certain topology given by KL divergence is a sufficient (though not a necessary) condition for the existence of a predictor. This is used to demonstrate that there is a solution to this problem for the set of all finite-memory process measures, complementing similar results obtained earlier in different settings. On the other hand, we show that there is no solution to this problem for the set of all stationary process measures, in contrast to a result of B. Ryabko (1988) that gives a solution to the

realizable case of this problem (that is, a predictor whose expected average KL error goes to zero if any stationary process is chosen to generate the data). Finally, we also consider a modified version of Problem 3, in which the performance of predictors is only compared on individual sequences. For this problem, we obtain, using a result from (Ryabko, 1986), a characterisation of those sets $\mathcal{C}$ for which a solution exists in terms of the Hausdorff dimension.

2. Notation and Definitions

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 $\Omega := (\mathcal{X}^\infty, \mathcal{B})$ where $\mathcal{B}$ is the sigma-field generated by the cylinder sets $[x_{1..n}]$, $x_i \in \mathcal{X}, n \in \mathbb{N}$ ($[x_{1..n}]$ 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

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

if $\mu(x_{1..n}) \neq 0$ and $\rho(x_{1..n}) \neq 0$, and $v(\mu, \rho, x_{1..n}) = 1$ otherwise.

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. Denote $\geq_{tv}$ the relation of absolute continuity (that is, $\rho \geq_{tv} \mu$ 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. Indeed, all these processes δ_p, $p \in [0, 1]$, 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).

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) cumulative (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.

We will use the following well-known identity (introduced, in the context of sequence prediction, by Ryabko, 1988)

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

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.

Definition 3 *Consider the following classes of process measures: $\mathcal{P}$ is the set of all process measures, $\mathcal{D}$ is the set of all degenerate discrete process measures, $\mathcal{S}$ is the set of all stationary processes and $\mathcal{M}_k$ is the set of all stationary measures with memory not greater than k (k-order Markov processes, with $\mathcal{M}_0$ being the set of all i.i.d. processes):*

$$\mathcal{D} := \{\mu \in \mathcal{P} : \exists x \in \mathcal{X}^\infty \;\; \mu(x) = 1\}, \tag{3}$$

$$\mathcal{S} := \{\mu \in \mathcal{P} : \forall n, k \geq 1 \, \forall a_{1..n} \in \mathcal{X}^n \, \mu(x_{1..n} = a_{1..n}) = \mu(x_{1+k..n+k} = a_{1..n})\}. \tag{4}$$

$$\begin{aligned} \mathcal{M}_k := \{\mu \in \mathcal{S} : \forall n \geq k \, \forall a \in \mathcal{X} \, \forall a_{1..n} \in \mathcal{X}^n \\ \mu(x_{n+1} = a | x_{1..n} = a_{1..n}) = \mu(x_{k+1} = a | x_{1..k} = a_{n-k+1..n})\}. \end{aligned} \tag{5}$$

Abusing the notation, we will sometimes use elements of $\mathcal{D}$ and $\mathcal{X}^\infty$ interchangeably. The following (simple and well-known) statement will be used repeatedly in the examples.

Lemma 4 *For every $\rho \in \mathcal{P}$ there exists $\mu \in \mathcal{D}$ such that $d_n(\mu, \rho) \geq n \log |\mathcal{X}|$ for all $n \in \mathbb{N}$.*

Proof Indeed, for each n we can select $\mu(x_n = a) = 1$ for $a \in \mathcal{X}$ that minimizes $\rho(x_n = a | x_{1..n-1})$, so that $\rho(x_{1..n}) \leq |\mathcal{X}|^{-n}$. ■

3. Sequence Prediction Problems

For the two notions of predictive quality introduced, we can now state formally the sequence prediction problems.

Problem 1(realizable case). Given a set of probability measures $\mathcal{C}$, find a measure ρ such that ρ predicts in total variation (expected average KL divergence) every $\mu \in \mathcal{C}$, if such a ρ exists.

Thus, Problem 1 is about finding a predictor for the case when the process generating the data is known to belong to a given class $\mathcal{C}$. That is, the set $\mathcal{C}$ here is a set of measures that generate the data. Next let us formulate the questions about $\mathcal{C}$ as a set of predictors.

Problem 2 (non-realizable case). Given a set of process measures (predictors) $\mathcal{C}$, find a process measure ρ such that ρ predicts in total variation (in expected average KL divergence) every measure $\nu \in \mathcal{P}$ such that there is $\mu \in \mathcal{C}$ which predicts (in the same sense) ν.

While Problem 2 is already quite general, it does not yet address what can be called the fully agnostic case: if nothing at all is known about the process ν generating the data, it means that there may be no $\mu \in \mathcal{C}$ such that μ predicts ν, and then, even if we have a solution ρ to the Problem 2, we still do not know what the performance of ρ is going to be on the data generated by ν, compared to the performance of the predictors from $\mathcal{C}$. To address this fully agnostic case we have to introduce the notion of loss.

Definition 5 *Introduce the almost sure total variation loss of ρ with respect to μ*

$$l_{tv}(\mu, \rho) := \inf\{\alpha \in [0,1] : \limsup_{n\to\infty} v(\mu, \rho, x_{1..n}) \leq \alpha \ \mu\text{–}a.s.\},$$

and the asymptotic KL loss

$$l_{KL}(\nu, \rho) := \limsup_{n\to\infty} \frac{1}{n} d_n(\nu, \rho).$$

We can now formulate the fully agnostic version of the sequence prediction problem.

Problem 3. Given a set of process measures (predictors) $\mathcal{C}$, find a process measure ρ such that ρ predicts at least as well as any μ in $\mathcal{C}$, if any process measure $\nu \in \mathcal{P}$ is chosen to generate the data:

$$l(\nu, \rho) - l(\nu, \mu) \leq 0 \tag{6}$$

for every $\nu \in \mathcal{P}$ and every $\mu \in \mathcal{C}$, where $l(\cdot, \cdot)$ is either $l_{tv}(\cdot, \cdot)$ or $l_{KL}(\cdot, \cdot)$.

The three problems just formulated represent different conceptual approaches to the sequence prediction problem. Let us illustrate the difference by the following **informal example**. Suppose that the set $\mathcal{C}$ is that of all (ergodic, finite-state) Markov chains. Markov chains being a familiar object in probability and statistics, we can easily construct a predictor ρ that predicts every $\mu \in \mathcal{C}$ (for example, in expected average KL divergence, see Krichevsky, 1993). That is, if we know that the process μ generating the data is Markovian, we know that our predictor is going to perform well. This is the realizable case of Problem 1. In reality, rarely can we be sure that the Markov assumption holds true for the data at hand. We may believe, however, that it is still a reasonable assumption, in the sense that there is a Markovian model which, for our purposes (for the purposes of prediction), is a good model of the data. Thus we may assume that there is a Markov model (a predictor) that predicts

well the process that we observe, and we would like to combine the predictive qualities of all these Markov models. This is the "non-realizable" case of Problem 2. Note that this problem is more difficult than the first one; in particular, a process ν generating the data may be singular with respect to any Markov process, and still be predicted well (in the sense of expected average KL divergence, for example) by some of them. Still, here we are making some assumptions about the process generating the data, and, if these assumptions are wrong, then we do not know anything about the performance of our predictor. Thus, we may ultimately wish to acknowledge that we do not know anything at all about the data; we still know a lot about Markov processes, and we would like to use this knowledge on our data. If there is anything at all Markovian in it (that is, anything that can be captured by a Markov model), then we would like our predictor to use it. In other words, we want to have a predictor that predicts any process measure whatsoever (at least) as well as any Markov predictor. This is the "fully agnostic" case of Problem 3.

Of course, Markov processes were just mentioned as an example, while in this work we are only concerned with the most general case of arbitrary (uncountable) sets $\mathcal{C}$ of process measures.

The following statement is rather obvious.

Proposition 6 *Any solution to Problem 3 is a solution to Problem 2, and any solution to Problem 2 is a solution to Problem 1.*

Despite the conceptual differences in formulations, it may be somewhat unclear whether the three problems are indeed different. It appears that this depends on the measure of predictive quality chosen: for the case of prediction in total variation distance all the three problems coincide, while for the case of prediction in expected average KL divergence they are different.

4. Prediction in Total Variation

As it was mentioned, a measure μ is absolutely continuous with respect to a measure ρ if and only if ρ predicts μ in total variation distance. This reduces studying at least Problem 1 for total variation distance to studying the relation of absolute continuity. Introduce the notation $\rho \geq_{tv} \mu$ for this relation.

Let us briefly recall some facts we know about $\geq_{tv}$; details can be found, for example, in (Plesner and Rokhlin, 1946). Let $[\mathcal{P}]_{tv}$ denote the set of equivalence classes of $\mathcal{P}$ with respect to $\geq_{tv}$, and for $\mu \in \mathcal{P}_{tv}$ denote $[\mu]$ the equivalence class that contains μ. Two elements $\sigma_1, \sigma_2 \in [\mathcal{P}]_{tv}$ (or $\sigma_1, \sigma_2 \in \mathcal{P}$) are called disjoint (or singular) if there is no $\nu \in [\mathcal{P}]_{tv}$ such that $\sigma_1 \geq_{tv} \nu$ and $\sigma_2 \geq_{tv} \nu$; in this case we write $\sigma_1 \perp_{tv} \sigma_2$. We write $[\mu_1] + [\mu_2]$ for $[\frac{1}{2}(\mu_1 + \mu_2)]$. Every pair $\sigma_1, \sigma_2 \in [\mathcal{P}]_{tv}$ has a supremum $\sup(\sigma_1, \sigma_2) = \sigma_1 + \sigma_2$. Introducing into $[\mathcal{P}]_{tv}$ an extra element 0 such that $\sigma \geq_{tv} 0$ for all $\sigma \in [\mathcal{P}]_{tv}$, we can state that for every $\rho, \mu \in [\mathcal{P}]_{tv}$ there exists a unique pair of elements μ_s and μ_a such that $\mu = \mu_a + \mu_s$, $\rho \geq \mu_a$ and $\rho \perp_{tv} \mu_s$. (This is a form of Lebesgue decomposition.) Moreover, $\mu_a = \inf(\rho, \mu)$. Thus, every pair of elements has a supremum and an infimum. Moreover, every bounded set of disjoint elements of $[\mathcal{P}]_{tv}$ is at most countable.

Furthermore, we introduce the (unconditional) total variation distance between process measures.

Definition 7 (unconditional total variation distance) *The (unconditional) total variation distance is defined as*

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

Known characterizations of those sets $\mathcal{C}$ that are bounded with respect to $\geq_{tv}$ can now be related to our prediction problems 1-3 as follows.

Theorem 8 *Let $\mathcal{C} \subset \mathcal{P}$. The following statements about $\mathcal{C}$ are equivalent.*

(i) There exists a solution to Problem 1 in total variation.

(ii) There exists a solution to Problem 2 in total variation.

(iii) There exists a solution to Problem 3 in total variation.

(iv) $\mathcal{C}$ is upper-bounded with respect to $\geq_{tv}$.

(v) There exists a sequence $\mu_k \in \mathcal{C}$, $k \in \mathbb{N}$ such that for some (equivalently, for every) sequence of weights $w_k \in (0,1]$, $k \in \mathbb{N}$ such that $\sum_{k\in\mathbb{N}} w_k = 1$, the measure $\nu = \sum_{k\in\mathbb{N}} w_k\mu_k$ satisfies $\nu \geq_{tv} \mu$ for every $\mu \in \mathcal{C}$.

(vi) $\mathcal{C}$ is separable with respect to the total variation distance.

(vii) Let $\mathcal{C}^+ := \{\mu \in \mathcal{P} : \exists \rho \in \mathcal{C}\ \rho \geq_{tv} \mu\}$. Every disjoint (with respect to $\geq_{tv}$) subset of $\mathcal{C}^+$ is at most countable.

Moreover, every solution to any of the Problems 1-3 is a solution to the other two, as is any upper bound for $\mathcal{C}$. The sequence μ_k in the statement (v) can be taken to be any dense (in the total variation distance) countable subset of $\mathcal{C}$ (cf. (vi)), or any maximal disjoint (with respect to $\geq_{tv}$) subset of $\mathcal{C}^+$ of statement (vii), in which every measure that is not in $\mathcal{C}$ is replaced by any measure from $\mathcal{C}$ that dominates it.

Proof The implications $(i) \Leftarrow (ii) \Leftarrow (iii)$ are obvious (cf. Proposition 6). The implication $(iv) \Rightarrow (i)$ is a reformulation of the result of Blackwell and Dubins (1962). The converse (and hence $(v) \Rightarrow (iv)$) was established in (Kalai and Lehrer, 1994). $(i) \Rightarrow (ii)$ follows from the equivalence $(i) \Leftrightarrow (iv)$ and the transitivity of $\geq_{tv}$; $(i) \Rightarrow (iii)$ follows from the transitivity of $\geq_{tv}$ and from Lemma 9 below: indeed, from Lemma 9 we have $l_{tv}(\nu,\mu) = 0$ if $\mu \geq_{tv} \nu$ and $l_{tv}(\nu,\mu) = 1$ otherwise. From this and the transitivity of $\geq_{tv}$ it follows that if $\rho \geq_{tv} \mu$ then also $l_{tv}(\nu,\rho) \leq l_{tv}(\nu,\mu)$ for all $\nu \in \mathcal{P}$. The equivalence of (v), (vi), and (i) was established in (Ryabko, 2010a). The equivalence of (iv) and (vii) was proven in (Plesner and Rokhlin, 1946). The concluding statements of the theorem are easy to demonstrate from the results cited above. ■

The following lemma is an easy consequence of (Blackwell and Dubins, 1962).

Lemma 9 *Let μ, ρ be two process measures. Then $v(\mu, \rho, x_{1..n})$ converges to either 0 or 1 with μ-probability 1.*

9

Proof Assume that μ is not absolutely continuous with respect to ρ (the other case is covered by (Blackwell and Dubins, 1962)). By Lebesgue decomposition theorem, the measure μ admits a representation $\mu = \alpha\mu_a + (1-\alpha)\mu_s$ where $\alpha \in [0,1]$ and the measures μ_a and μ_s are such that μ_a is absolutely continuous with respect to ρ and μ_s is singular with respect to ρ. Let W be such a set that $\mu_a(W) = \rho(W) = 1$ and $\mu_s(W) = 0$. Note that we can take $\mu_a = \mu|_W$ and $\mu_s = \mu|_{\mathcal{X}^\infty \setminus W}$. From (Blackwell and Dubins, 1962) we have $v(\mu_a, \rho, x_{1..n}) \to 0$ μ_a-a.s., as well as $v(\mu_a, \mu, x_{1..n}) \to 0$ μ_a-a.s. and $v(\mu_s, \mu, x_{1..n}) \to 0$ μ_s-a.s. Moreover, $v(\mu_s, \rho, x_{1..n}) \geq |\mu_s(W|x_{1..n}) - \rho(W|x_{1..n})| = 1$ so that $v(\mu_s, \rho, x_{1..n}) \to 1$ μ_s-a.s. Furthermore,

$$v(\mu, \rho, x_{1..n}) \leq v(\mu, \mu_a, x_{1..n}) + v(\mu_a, \rho, x_{1..n}) = I$$

and

$$v(\mu, \rho, x_{1..n}) \geq -v(\mu, \mu_s, x_{1..n}) + v(\mu_s, \rho, x_{1..n}) = II.$$

We have $I \to 0$ μ_a-a.s. and hence $\mu|_W$-a.s., as well as $II \to 1$ μ_s-a.s. and hence $\mu|_{\mathcal{X}^\infty \setminus W}$-a.s. Thus, $\mu(v(\mu, \rho, x_{1..n}) \to 0 \text{ or } 1) \leq \mu(W)\mu|_W(I \to 0) + \mu(\mathcal{X}^\infty \setminus W)\mu|_{\mathcal{X}^\infty \setminus W}(II \to 1) = \mu(W) + \mu(\mathcal{X}^\infty \setminus W) = 1$, which concludes the proof. ■

Remark. Using Lemma 9 we can also define *expected* (rather than almost sure) total variation loss of ρ with respect to μ, as the μ-probability that $v(\mu, \rho)$ converges to 1:

$$l'_{tv}(\mu, \rho) := \mu\{x_1, x_2, \dots \in \mathcal{X}^\infty : v(\mu, \rho, x_{1..n}) \to 1\}.$$

Then Problem 3 can be reformulated for this notion of loss. However, it is easy to see that for this reformulation Theorem 8 holds true as well.

Thus, we can see that, for the case of prediction in total variation, all the sequence prediction problems formulated reduce to studying the relation of absolute continuity for process measures and those families of measures that are absolutely continuous (have a density) with respect to some measure (a predictor). On the one hand, from a 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. (In particular, this means that they are large from a computational point of view.) 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 (as explained in Section 2).

That is why we have to consider weaker notions of predictions; from these, prediction in expected average KL divergence is perhaps one of the weakest. The goal of the next sections is to see which of the properties that we have for total variation can be transferred (and in which sense) to the case of expected average KL divergence.

5. Prediction in Expected Average KL Divergence

First of all, we have to observe that for prediction in KL divergence Problems 1, 2, and 3 are different, as the following theorem shows. While the examples provided in the proof are

artificial, there is a very important example illustrating the difference between Problem 1 and Problem 3 for expected average KL divergence: the set $\mathcal{S}$ of all stationary processes, given in Theorem 16 in the end of this section.

Theorem 10 *For the case of prediction in expected average KL divergence, Problems 1, 2 and 3 are different: there exists a set $\mathcal{C}_1 \subset \mathcal{P}$ for which there is a solution to Problem 1 but there is no solution to Problem 2, and there is a set $\mathcal{C}_2 \subset \mathcal{P}$ for which there is a solution to Problem 2 but there is no solution to Problem 3.*

Proof We have to provide two examples. Fix the binary alphabet $\mathcal{X} = \{0, 1\}$. For each deterministic sequence $t = t_1, t_2, \cdots \in \mathcal{X}^\infty$ construct the process measure γ_t as follows: $\gamma_t(x_n = t_n | t_{1..n-1}) := 1 - \frac{1}{n+1}$ and for $x_{1..n-1} \neq t_{1..n-1}$ let $\gamma_t(x_n = 0 | x_{1..n-1}) = 1/2$, for all $n \in \mathbb{N}$. That is, γ_t is Bernoulli i.i.d. $1/2$ process measure strongly biased towards a specific deterministic sequence, t. Let also $\gamma(x_{1..n}) = 2^{-n}$ for all $x_{1..n} \in \mathcal{X}^n$, $n \in \mathbb{N}$ (the Bernoulli i.i.d. $1/2$). For the set $\mathcal{C}_1 := \{\gamma_t : t \in \mathcal{X}^\infty\}$ we have a solution to Problem 1: indeed, $d_n(\gamma_t, \gamma) \leq 1 = o(n)$. However, there is no solution to Problem 2. Indeed, for each $t \in \mathcal{D}$ we have $d_n(t, \gamma_t) = \log n = o(n)$ (that is, for every deterministic measure there is an element of $\mathcal{C}_1$ which predicts it), while by Lemma 4 for every $\rho \in \mathcal{P}$ there exists $t \in \mathcal{D}$ such that $d_n(t, \rho) \geq n$ for all $n \in \mathbb{N}$ (that is, there is no predictor which predicts every measure that is predicted by at least one element of $\mathcal{C}_1$).

The second example is similar. For each deterministic sequence $t = t_1, t_2, \cdots \in \mathcal{D}$ construct the process measure γ_t as follows: $\gamma'_t(x_n = t_n | t_{1..n-1}) := 2/3$ and for $x_{1..n-1} \neq t_{1..n-1}$ let $\gamma'_t(x_n = 0 | x_{1..n-1}) = 1/2$, for all $n \in \mathbb{N}$. It is easy to see that γ is a solution to Problem 2 for the set $\mathcal{C}_2 := \{\gamma'_t : t \in \mathcal{X}^\infty\}$. Indeed, if $\nu \in \mathcal{P}$ is such that $d_n(\nu, \gamma') = o(n)$ then we must have $\nu(t_{1..n}) = o(1)$. From this and the fact that γ and γ' coincide (up to $O(1)$) on all other sequences we conclude $d_n(\nu, \gamma) = o(n)$. However, there is no solution to Problem 3 for $\mathcal{C}_2$. Indeed, for every $t \in \mathcal{D}$ we have $d_n(t, \gamma'_t) = n \log 3/2 + o(n)$. Therefore, if ρ is a solution to Problem 3 then $\limsup \frac{1}{n} d_n(t, \rho) \leq \log 3/2 < 1$ which contradicts Lemma 4. ■

Thus, prediction in expected average KL divergence turns out to be a more complicated matter than prediction in total variation. The next idea is to try and see which of the facts about prediction in total variation can be generalized to some of the problems concerning prediction in expected average KL divergence.

First, observe that, for the case of prediction in total variation, the equivalence of Problems 1 and 2 was derived from the transitivity of the relation $\geq_{tv}$ of absolute continuity. For the case of expected average KL divergence, the relation "ρ predicts μ in expected average KL divergence" is not transitive (and Problems 1 and 2 are not equivalent). However, for Problem 2 we are interested in the following relation: ρ "dominates" μ if ρ predicts every ν such that μ predicts ν. Denote this relation by $\geq_{KL}$:

Definition 11 ($\geq_{KL}$) *We write $\rho \geq_{KL} \mu$ if for every $\nu \in \mathcal{P}$ the equality $\limsup \frac{1}{n} d_n(\nu, \mu) = 0$ implies $\limsup \frac{1}{n} d_n(\nu, \rho) = 0$.*

The relation $\geq_{KL}$ has some similarities with $\geq_{tv}$. First of all, $\geq_{KL}$ is also transitive (as can be easily seen from the definition). Moreover, similarly to $\geq_{tv}$, one can show that for any μ, ρ any strictly convex combination $\alpha\mu + (1 - \alpha)\rho$ is a supremum of $\{\rho, \mu\}$ with respect

to $\geq_{KL}$. Next we will obtain a characterization of predictability with respect to $\geq_{KL}$ similar to one of those obtained for $\geq_{tv}$.

The key observation is the following. If there is a solution to Problem 2 for a set $\mathcal{C}$ then a solution can be obtained as a Bayesian mixture over a countable subset of $\mathcal{C}$. For total variation this is the statement (v) of Theorem 8.

Theorem 12 *Let $\mathcal{C}$ be a set of probability measures on Ω. If there is a measure ρ such that $\rho \geq_{KL} \mu$ for every $\mu \in \mathcal{C}$ (ρ is a solution to Problem 2) then there is a sequence $\mu_k \in \mathcal{C}$, $k \in \mathbb{N}$, such that $\sum_{k\in\mathbb{N}} w_k\mu_k \geq_{KL} \mu$ for every $\mu \in \mathcal{C}$, where w_k are some positive weights.*

The proof is deferred to Section 7. An analogous result for Problem 1 was established in (Ryabko, 2009). (The proof of Theorem 12 is based on similar ideas, but is more involved.)

For the case of Problem 3, we do not have results similar to Theorem 12 (or statement (v) of Theorem 8); in fact, we conjecture that the opposite is true: there exists a (measurable) set $\mathcal{C}$ of measures such that there is a solution to Problem 3 for $\mathcal{C}$, but there is no Bayesian solution to Problem 3, meaning that there is no probability distribution on $\mathcal{C}$ (discrete or not) such that the mixture over $\mathcal{C}$ with respect to this distribution is a solution to Problem 3 for $\mathcal{C}$.

However, we can take a different route and extend another part of Theorem 8 to obtain a characterization of sets $\mathcal{C}$ for which a solution to Problem 3 exists.

We have seen that, in the case of prediction in total variation, separability with respect to the topology of this distance is a necessary and sufficient condition for the existence of a solution to Problems 1-3. In the case of expected average KL divergence the situation is somewhat different, since, first of all, (asymptotic average) KL divergence is not a metric. 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 the next theorem.

Definition 13 *Define the distance $d_\infty(\mu_1, \mu_2)$ on process measures as follows*

$$d_\infty(\mu_1, \mu_2) = \limsup_{n\to\infty} \sup_{x_{1..n}\in\mathcal{X}^n} \frac{1}{n} \left| \log \frac{\mu_1(x_{1..n})}{\mu_2(x_{1..n})} \right|, \tag{7}$$

where we assume $\log 0/0 := 0$.

Clearly, d_∞ is symmetric and satisfies the triangle inequality, but it is not exact. Moreover, for every μ_1, μ_2 we have

$$\limsup_{n\to\infty} \frac{1}{n} d_n(\mu_1, \mu_2) \leq d_\infty(\mu_1, \mu_2). \tag{8}$$

The distance $d_\infty(\mu_1, \mu_2)$ measures the difference in behaviour of μ_1 and μ_2 on all individual sequences. Thus, using this distance to analyse Problem 3 is most close to the traditional approach to the non-realizable case, which is formulated in terms of predicting individual deterministic sequences.

Theorem 14 *(i) Let $\mathcal{C}$ be a set of process measures. If $\mathcal{C}$ is separable with respect to d_∞ then there is a solution to Problem 3 for $\mathcal{C}$, for the case of prediction in expected average KL divergence.*

(ii) There exists a set of process measures $\mathcal{C}$ such that $\mathcal{C}$ is not separable with respect to d_∞, but there is a solution to Problem 3 for this set, for the case of prediction in expected average KL divergence.

Proof For the first statement, let $\mathcal{C}$ be separable and let $(\mu_k)_{k\in\mathbb{N}}$ be a dense countable subset of $\mathcal{C}$. Define $\nu := \sum_{k\in\mathbb{N}} w_k\mu_k$, where w_k are any positive summable weights. Fix any measure τ and any $\mu \in \mathcal{C}$. We will show that $\limsup_{n\to\infty} \frac{1}{n} d_n(\tau,\nu) \le \limsup_{n\to\infty} \frac{1}{n} d_n(\tau,\mu)$. For every ε, find such a $k \in \mathbb{N}$ that $d_\infty(\mu,\mu_k) \le \varepsilon$. We have

$$\begin{aligned} d_n(\tau,\nu) \le d_n(\tau, w_k\mu_k) &= \mathbf{E}_\tau \log \frac{\tau(x_{1..n})}{\mu_k(x_{1..n})} - \log w_k \\ &= \mathbf{E}_\tau \log \frac{\tau(x_{1..n})}{\mu(x_{1..n})} + \mathbf{E}_\tau \log \frac{\mu(x_{1..n})}{\mu_k(x_{1..n})} - \log w_k \\ &\le d_n(\tau,\mu) + \sup_{x_{1..n}\in\mathcal{X}^n} \log \left| \frac{\mu(x_{1..n})}{\mu_k(x_{1..n})} \right| - \log w_k. \end{aligned}$$

From this, dividing by n taking $\limsup_{n\to\infty}$ on both sides, we conclude

$$\limsup_{n\to\infty} \frac{1}{n} d_n(\tau,\nu) \le \limsup_{n\to\infty} \frac{1}{n} d_n(\tau,\mu) + \varepsilon.$$

Since this holds for every $\varepsilon > 0$ the first statement is proven.

The second statement is proven by the following example. 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 $\sqrt{n}$, for all $n \in \mathbb{N}$. Clearly, this set is uncountable. It is easy to check that $\mu_1 \ne \mu_2$ implies $d_\infty(\mu_1,\mu_2) = \infty$ 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. Since all elements of $\mathcal{C}$ are deterministic, ν is also a solution to Problem 3 for $\mathcal{C}$. ■

Although simple, Theorem 14 can be used to establish the existence of a solution to Problem 3 for an important class of process measures: that of all processes with finite memory, as the next theorem shows. Results similar to Theorem 15 are known in different settings, e.g., (Ziv and Lempel, 1978; Ryabko, 1984; Cesa-Bianchi and Lugosi, 1999) and others.

Theorem 15 *There exists a solution to Problem 3 for prediction in expected average KL divergence for the set of all finite-memory process measures $\mathcal{M} := \cup_{k\in\mathbb{N}}\mathcal{M}_k$.*

Proof We will show that the set $\mathcal{M}$ is separable with respect to d_∞. Then the statement will follow from Theorem 14. It is enough to show that each set $\mathcal{M}_k$ is separable with respect to d_∞.

For simplicity, assume that the alphabet is binary ($|\mathcal{X}| = 2$; the general case is analogous). Observe that the family $\mathcal{M}_k$ of k-order stationary binary-valued Markov processes is parametrized by $|\mathcal{X}|^k$ $[0,1]$-valued parameters: probability of observing 0 after observing $x_{1..k}$, for each $x_{1..k} \in \mathcal{X}^k$. Note that this parametrization is continuous (as a mapping from

the parameter space with the Euclidean topology to $\mathcal{M}_k$ with the topology of d_∞). Indeed, for any $\mu_1, \mu_2 \in \mathcal{M}_k$ and every $x_{1..n} \in \mathcal{X}^n$ such that $\mu_i(x_{1..n}) \neq 0$, $i = 1, 2$, it is easy to see that

$$\frac{1}{n}\left|\log \frac{\mu_1(x_{1..n})}{\mu_2(x_{1..n})}\right| \leq \sup_{x_{1..k+1}} \frac{1}{k+1}\left|\log \frac{\mu_1(x_{1..k+1})}{\mu_2(x_{1..k+1})}\right|, \tag{9}$$

so that the right-hand side of (9) also upper-bounds $d_\infty(\mu_1, \mu_2)$, implying continuity of the parametrization.

It follows that the set μ_q^k, $q \in Q^{|\mathcal{X}|^k}$ of all stationary k-order Markov processes with rational values of all the parameters ($Q := \mathbb{Q} \cap [0, 1]$) is dense in $\mathcal{M}_k$, proving the separability of the latter set. ■

Another important example is the set of all stationary process measures $\mathcal{S}$. This example also illustrates the difference between the prediction problems that we consider. For this set a solution to Problem 1 was given in (Ryabko, 1988). In contrast, here we show that there is no solution to Problem 3 for $\mathcal{S}$.

Theorem 16 *There is no solution to Problem 3 for the set of all stationary processes $\mathcal{S}$.*

Proof This proof is based on the construction similar to the one used in (Ryabko, 1988) to demonstrate impossibility of consistent prediction of stationary processes without Cesaro averaging.

Let m be a Markov chain with states $0, 1, 2, \ldots$ and state transitions defined as follows. From each sate $k \in \mathbb{N} \cup \{0\}$ the chain passes to the state $k + 1$ with probability $2/3$ and to the state 0 with probability $1/3$. It is easy to see that this chain possesses a unique stationary distribution on the set of states (see, e.g., Shiryaev, 1996); taken as the initial distribution it defines a stationary ergodic process with values in $\mathbb{N} \cup \{0\}$. Fix the ternary alphabet $\mathcal{X} = \{a, 0, 1\}$. For each sequence $t = t_1, t_2, \cdots \in \{0, 1\}^\infty$ define the process μ_t as follows. It is a deterministic function of the chain m. If the chain is in the state 0 then the process μ_t outputs a; if the chain m is in the state $k > 0$ then the process outputs t_k. That is, we have defined a hidden Markov process which in the state 0 of the underlying Markov chain always outputs a, while in other states it outputs either 0 or 1 according to the sequence t.

To show that there is no solution to Problem 3 for $\mathcal{S}$, we will show that there is no solution to Problem 3 for the smaller set $\mathcal{C} := \{\mu_t : t \in \{0, 1\}^\infty\}$. Indeed, for any $t \in \{0, 1\}^\infty$ we have $d_n(t, \mu_t) = n \log 3/2 + o(n)$. Then if ρ is a solution to Problem 3 for $\mathcal{C}$ we should have $\limsup_{n\to\infty} \frac{1}{n} d_n(t, \rho) \leq \log 3/2 < 1$ for every $t \in \mathcal{D}$, which contradicts Lemma 4. ■

From the proof of Theorem 16 one can see that, in fact, the statement that is proven is stronger: there is no solution to Problem 3 for the set of all functions of stationary ergodic countable-state Markov chains. We conjecture that a solution to Problem 2 exists for the latter set, but not for the set of all stationary processes.

As we have seen in the statements above, the set of all deterministic measures $\mathcal{D}$ plays an important role in the analysis of the predictors in the sense of Problem 3. Therefore, an interesting question is to characterize those sets $\mathcal{C}$ of measures for which there is a predictor ρ that predicts *every individual sequence* at least as well as any measure from $\mathcal{C}$. Such a

characterization can be obtained in terms of Hausdorff dimension, using a result of Ryabko (1986), that shows that Hausdorff dimension of a set characterizes the optimal prediction error that can be attained by any predictor.

For a set $A \subset \mathcal{X}^\infty$ denote $H(A)$ its Hausdorff dimension (see, for example, (Billingsley, 1965) for its definition).

Theorem 17 *Let $\mathcal{C} \subset \mathcal{P}$. The following statements are equivalent.*

(i) There is a measure $\rho \in \mathcal{P}$ that predicts every individual sequence at least as well as the best measure from $\mathcal{C}$: for every $\mu \in \mathcal{C}$ and every sequence $x_1, x_2, \dots \in \mathcal{X}^\infty$ we have

$$\liminf_{n\to\infty} -\frac{1}{n}\log\rho(x_{1..n}) \le \liminf_{n\to\infty} -\frac{1}{n}\log\mu(x_{1..n}). \tag{10}$$

(ii) For every $\alpha \in [0,1]$ the Hausdorff dimension of the set of sequences on which the average prediction error of the best measure in $\mathcal{C}$ is not greater than α is bounded by $\alpha/\log|\mathcal{X}|$:

$$H(\{x_1, x_2, \dots \in \mathcal{X}^\infty : \inf_{\mu\in\mathcal{C}} \liminf_{n\to\infty} -\frac{1}{n}\log\mu(x_{1..n}) \le \alpha\}) \le \alpha/\log|\mathcal{X}|. \tag{11}$$

Proof The implication $(i) \Rightarrow (ii)$ follows directly from (Ryabko, 1986) where it is shown that for every measure ρ one must have $H(\{x_1, x_2, \dots \in \mathcal{X}^\infty : \liminf_{n\to\infty} -\frac{1}{n}\log\rho(x_{1..n}) \le \alpha\}) \le \alpha/\log|\mathcal{X}|$.

To show the opposite implication, we again refer to (Ryabko, 1986): for every set $A \subset \mathcal{X}^\infty$ there is a measure ρ_A such that

$$\liminf_{n\to\infty} -\frac{1}{n}\log\rho_A(x_{1..n}) \le H(A)\log|\mathcal{X}|. \tag{12}$$

For each $\alpha \in [0,1]$ define $A_\alpha := \{x_1, x_2, \dots \in \mathcal{X}^\infty : \inf_{\mu\in\mathcal{C}} \liminf_{n\to\infty} -\frac{1}{n}\log\mu(x_{1..n}) \le \alpha\})$. By assumption, $H(A_\alpha) \le \alpha/\log|X|$, so that from (12) for all $x_1, x_2, \dots \in A_\alpha$ we obtain

$$\liminf_{n\to\infty} -\frac{1}{n}\log\rho_A(x_{1..n}) \le \alpha. \tag{13}$$

Furthermore, define $\rho := \sum_{q\in Q} w_q \rho_{A_q}$, where $Q = [0,1]\cap\mathbb{Q}$ is the set of rationals in $[0,1]$ and $(w_q)_{q\in Q}$ is any sequence of positive reals satisfying $\sum_{q\in Q} w_q = 1$. For every $\alpha \in [0,1]$ let $q_k \in Q$, $k \in \mathbb{N}$ be such a sequence that $0 \le q_k - \alpha \le 1/k$. Then, for every $n \in \mathbb{N}$ and every $x_1, x_2, \dots \in A_{q_k}$ we have

$$-\frac{1}{n}\log\rho(x_{1..n}) \le -\frac{1}{n}\log\rho_q(x_{1..n}) - \frac{\log w_{q_k}}{n}.$$

From this and (13) we get

$$\liminf_{n\to\infty} -\frac{1}{n}\log\rho(x_{1..n}) \le \liminf_{n\to\infty} \rho_{q_k}(x_{1..n}) + 1/k \le q_k + 1/k.$$

Since this holds for every $k \in \mathbb{N}$, it follows that for all $x_1, x_2, \dots \in \cap_{k\in\mathbb{N}} A_{q_k} = A_\alpha$ we have

$$\liminf_{n\to\infty} -\frac{1}{n}\log\rho(x_{1..n}) \le \inf_{k\in\mathbb{N}}(q_k + 1/k) = \alpha,$$

which completes the proof of the implication $(ii) \Rightarrow (i)$. ■

6. Discussion

It has been long realized that the so-called probabilistic and agnostic (adversarial, non-stochastic, deterministic) settings of the problem of sequential prediction are strongly related. This has been most evident from looking at the solutions to these problems, which are usually based on the same ideas. Here we have proposed a formulation of the agnostic problem as a non-realizable case of the probabilistic problem. While being very close to the traditional one, this setting allows us to directly compare the two problems. As a somewhat surprising result, we can see that whether the two problems are different depends on the measure of performance chosen: in the case of prediction in total variation distance they coincide, while in the case of prediction in expected average KL divergence they are different. In the latter case, the distinction becomes particularly apparent on the example of stationary processes: while a solution to the realizable problem has long been known, here we have shown that there is no solution to the agnostic version of this problem. This formalization also allowed us to introduce another problem that lies in between the realizable and the fully agnostic problems: given a class of process measures $\mathcal{C}$, find a predictor whose predictions are asymptotically correct for every measure for which at least one of the measures in $\mathcal{C}$ gives asymptotically correct predictions (Problem 2). This problem is less restrictive then the fully agnostic one (in particular, it is not concerned with the behaviour of a predictor on every deterministic sequence) but at the same time the solutions to this problem have performance guarantees far outside the model class considered.

In essence, the formulation of Problem 2 suggests to assume that we have a set of models one of which is good enough to make predictions, with the goal of combining the predictive powers of these models. This is perhaps a good compromise between making modelling assumptions on the data (the data is generated by one of the models we have) and the fully agnostic, worst-case, setting.

Since the problem formulations presented here are mostly new (at least, in such a general form), it is not surprising that there are many questions left open. A promising route to obtain new results seems to be to first analyse the case of prediction in total variation, which amounts to studying the relation of absolute continuity and singularity of probability measures, and then to try and find analogues in less restrictive (and thus more interesting and difficult) cases of predicting only the next observation, possibly with Cesaro averaging. This is the approach that we took in this work. Here it is interesting to find properties common to all or most of the prediction problems (in total variation as well as with respect to other measures of the performance), if it is at all possible. For example, the "countable Bayes" property of Theorem 12, that states that if there is a solution to a given sequence prediction problem for a set $\mathcal{C}$ then a solution can be obtained as a mixture over a suitable countable subset of $\mathcal{C}$, holds for Problems 1–3 in total variation, and for Problems 1 and 2 in KL divergence; however we conjecture that it does not hold for the Problem 3 in KL divergence.

It may also be interesting to study algebraic properties of the relation $\geq_{KL}$ that arises when studying Problem 2. We have show that it shares some properties with the relation $\geq_{tv}$ of absolute continuity. Since the latter characterizes prediction in total variation and the former characterizes prediction in KL divergence (in the sense of Problem 2), which is much weaker, it would be interesting to see exactly what properties the two relations share.

Another direction for future research concerns finite-time performance analysis. In this work we have adopted the asymptotic approach to the prediction problem, ignoring the behaviour of predictors before asymptotic. While for prediction in total variation it is a natural choice, for other measures of performance, including average KL divergence, it is clear that Problems 1-3 admit non-asymptotic formulations. It is also interesting what are the relations between performance guarantees that can be obtained in non-asymptotic formulations of Problems 1–3.

7. Proof of Theorem 12

Proof Define the sets C_μ as the set of all measures $\tau \in \mathcal{P}$ such that μ predicts τ in expected average KL divergence. Let $\mathcal{C}^+ := \cup_{\mu\in\mathcal{C}} C_\mu$. For each $\tau \in \mathcal{C}^+$ let $p(\tau)$ be any (fixed) $\mu \in \mathcal{C}$ such that $\tau \in C_\mu$. In other words, $\mathcal{C}^+$ is the set of all measures that are predicted by some of the measures in $\mathcal{C}$, and for each measure τ in $\mathcal{C}^+$ we designate one "parent" measure $p(\tau)$ from $\mathcal{C}$ such that $p(\tau)$ predicts τ.

Define the weights $w_k := 1/k(k+1)$, for all $k \in \mathbb{N}$.
Step 1. For each $\mu \in \mathcal{C}^+$ let δ_n be any monotonically increasing function such that $\delta_n(\mu) = o(n)$ and $d_n(\mu, p(\mu)) = o(\delta_n(\mu))$. Define the sets

$$U^n_\mu := \left\{ x_{1..n} \in \mathcal{X}^n : \mu(x_{1..n}) \geq \frac{1}{n}\rho(x_{1..n}) \right\}, \tag{14}$$

$$V^n_\mu := \left\{ x_{1..n} \in \mathcal{X}^n : p(\mu)(x_{1..n}) \geq 2^{-\delta_n(\mu)}\mu(x_{1..n}) \right\}, \tag{15}$$

and

$$T^n_\mu := U^n_\mu \cap V^n_\mu. \tag{16}$$

We will upper-bound $\mu(T^n_\mu)$. First, using Markov's inequality, we derive

$$\mu(\mathcal{X}^n \backslash U^n_\mu) = \mu\left(\frac{\rho(x_{1..n})}{\mu(x_{1..n})} > n\right) \leq \frac{1}{n} E_\mu \frac{\rho(x_{1..n})}{\mu(x_{1..n})} = \frac{1}{n}. \tag{17}$$

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

$$\begin{aligned} -\sum_{x_{1..n}\in A} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} &= -\mu(A) \sum_{x_{1..n}\in A} \frac{1}{\mu(A)} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &\geq -\mu(A) \log \frac{\rho(A)}{\mu(A)} \geq -\mu(A) \log \rho(A) - \frac{1}{2}. \end{aligned} \tag{18}$$

Moreover,

$$\begin{aligned} d_n(\mu, p(\mu)) = &- \sum_{x_{1..n}\in\mathcal{X}^n\backslash V^n_\mu} \mu(x_{1..n}) \log \frac{p(\mu)(x_{1..n})}{\mu(x_{1..n})} \\ &- \sum_{x_{1..n}\in V^n_\mu} \mu(x_{1..n}) \log \frac{p(\mu)(x_{1..n})}{\mu(x_{1..n})} \geq \delta_n(\mu_n)\mu(\mathcal{X}^n\backslash V^n_\mu) - 1/2, \end{aligned}$$

where in the inequality we have used (15) for the first summand and (18) for the second. Thus,

$$\mu(\mathcal{X}^n\backslash V_\mu^n) \le \frac{d_n(\mu, p(\mu)) + 1/2}{\delta_n(\mu)} = o(1). \tag{19}$$

From (16), (17) and (19) we conclude

$$\mu(\mathcal{X}^n\backslash T_\mu^n) \le \mu(\mathcal{X}^n\backslash V_\mu^n) + \mu(\mathcal{X}^n\backslash U_\mu^n) = o(1). \tag{20}$$

Step 2n: a countable cover, time n. Fix an $n \in \mathbb{N}$. Define $m_1^n := \max_{\mu\in\mathcal{C}} \rho(T_\mu^n)$ (since $\mathcal{X}^n$ are finite all suprema are reached). Find any μ_1^n such that $\rho_1^n(T_{\mu_1^n}^n) = m_1^n$ and let $T_1^n := T_{\mu_1^n}^n$. For $k > 1$, let $m_k^n := \max_{\mu\in\mathcal{C}} \rho(T_\mu^n\backslash T_{k-1}^n)$. If $m_k^n > 0$, let μ_k^n be any $\mu \in \mathcal{C}$ such that $\rho(T_{\mu_k^n}^n\backslash T_{k-1}^n) = m_k^n$, and let $T_k^n := T_{k-1}^n \cup T_{\mu_k^n}^n$; otherwise let $T_k^n := T_{k-1}^n$. 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

$$\nu_n := \sum_{k=1}^{K_n} w_k p(\mu_k^n). \tag{21}$$

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 the definitions (16), (14) and (15) we obtain

$$\nu_n(x_{1..n}) \ge w_k \frac{1}{n} 2^{-\delta_n(\mu)} \rho(x_{1..n}). \tag{22}$$

Step 2: the resulting predictor. Finally, define

$$\nu := \frac{1}{2}\gamma + \frac{1}{2}\sum_{n\in\mathbb{N}} w_n \nu_n, \tag{23}$$

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,

$$\rho(T_\mu^n\backslash T_{j_\mu^n}^n) \le \varepsilon_\mu^n. \tag{24}$$

We can upper-bound $\mu(T^n_\mu \backslash T^n_{j^n_\mu})$ as follows. First, observe that

$$\begin{aligned} d_n(\mu,\rho) = &- \sum_{x_{1..n}\in T^n_\mu \cap T^n_{j^n_\mu}} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &- \sum_{x_{1..n}\in T^n_\mu \backslash T^n_{j^n_\mu}} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &- \sum_{x_{1..n}\in \mathcal{X}^n \backslash T^n_\mu} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &= I + II + III. \quad (25) \end{aligned}$$

Then, from (16) and (14) we get

$$I \geq -\log n. \quad (26)$$

From (18) and (24) we get

$$II \geq -\mu(T^n_\mu \backslash T^n_{j^n_\mu}) \log \rho(T^n_\mu \backslash T^n_{j^n_\mu}) - 1/2 \geq -\mu(T^n_\mu \backslash T^n_{j^n_\mu}) \log \varepsilon^n_\mu - 1/2. \quad (27)$$

Furthermore,

$$\begin{aligned} III \geq &\sum_{x_{1..n}\in \mathcal{X}^n \backslash T^n_\mu} \mu(x_{1..n}) \log \mu(x_{1..n}) \\ &\geq \mu(\mathcal{X}^n \backslash T^n_\mu) \log \frac{\mu(\mathcal{X}^n \backslash T^n_\mu)}{|\mathcal{X}^n \backslash T^n_\mu|} \geq -\frac{1}{2} - \mu(\mathcal{X}^n \backslash T^n_\mu) n \log |\mathcal{X}|, \quad (28) \end{aligned}$$

where the first inequality is obvious, in the second inequality we have used the fact that entropy is maximized when all events are equiprobable and in the third one we used $|\mathcal{X}^n \backslash T^n_\mu| \leq |\mathcal{X}|^n$. Combining (25) with the bounds (26), (27) and (28) we obtain

$$d_n(\mu,\rho) \geq -\log n - \mu(T^n_\mu \backslash T^n_{j^n_\mu}) \log \varepsilon^n_\mu - 1 - \mu(\mathcal{X}^n \backslash T^n_\mu) n \log |\mathcal{X}|,$$

so that

$$\mu(T^n_\mu \backslash T^n_{j^n_\mu}) \leq \frac{1}{-\log \varepsilon^n_\mu} \Big(d_n(\mu,\rho) + \log n + 1 + \mu(\mathcal{X}^n \backslash T^n_\mu) n \log |\mathcal{X}| \Big). \quad (29)$$

From the fact that $d_n(\mu,\rho) = o(n)$ and (20) it follows that the term in brackets is $o(n)$, so that we can define the parameters ε^n_μ in such a way that $-\log \varepsilon^n_\mu = o(n)$ while at the same time the bound (29) gives $\mu(T^n_\mu \backslash T^n_{j^n_\mu}) = o(1)$. Fix such a choice of ε^n_μ. Then, using (20), we conclude

$$\mu(\mathcal{X}^n \backslash T^n_{j^n_\mu}) \leq \mu(\mathcal{X}^n \backslash T^n_\mu) + \mu(T^n_\mu \backslash T^n_{j^n_\mu}) = o(1). \quad (30)$$

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

$$\nu(x_{1..n}) \geq \frac{1}{2} w_n \nu_n(x_{1..n}) \geq \frac{1}{2} w_n w_{j^n_\mu} \frac{1}{n} 2^{-\delta_n(\mu)} \rho(x_{1..n}) \geq \frac{w_n}{4n} (\varepsilon^n_\mu)^2 2^{-\delta_n(\mu)} \rho(x_{1..n}), \quad (31)$$

where the first inequality follows from (23), the second from (22), and in the third we have used $w_{j_\mu^n} = 1/(j_\mu^n)(j_\mu^n+1)$ and $j_\mu^n = 1/\varepsilon_n^\mu$. Next we use the decomposition

$$d_n(\mu,\nu) = -\sum_{x_{1..n}\in T^n_{j_\mu^n}} \mu(x_{1..n}) \log \frac{\nu(x_{1..n})}{\mu(x_{1..n})} - \sum_{x_{1..n}\in \mathcal{X}^n\setminus T^n_{j_\mu^n}} \mu(x_{1..n}) \log \frac{\nu(x_{1..n})}{\mu(x_{1..n})} = I + II. \quad (32)$$

From (31) we find

$$\begin{aligned} I &\le -\log\Big(\frac{w_n}{4n}(\varepsilon_\mu^n)^2 2^{-\delta_n(\mu)}\Big) - \sum_{x_{1..n}\in T^n_{j_\mu^n}} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &= (o(n) - 2\log\varepsilon_\mu^n + \delta_n(\mu)) + \left(d_n(\mu,\rho) + \sum_{x_{1..n}\in \mathcal{X}^n\setminus T^n_{j_\mu^n}} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \right) \\ &\le o(n) - \sum_{x_{1..n}\in \mathcal{X}^n\setminus T^n_{j_\mu^n}} \mu(x_{1..n}) \log \mu(x_{1..n}) \\ &\le o(n) + \mu(\mathcal{X}^n\setminus T^n_{j_\mu^n}) n \log|\mathcal{X}| = o(n), \quad (33) \end{aligned}$$

where in the second inequality we have used $-\log\varepsilon_\mu^n = o(n)$, $d_n(\mu,\rho) = o(n)$ and $\delta_n(\mu) = o(n)$, 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 (30). Moreover, from (23) we find

$$II \le \log 2 - \sum_{x_{1..n}\in \mathcal{X}^n\setminus T^n_{j_\mu^n}} \mu(x_{1..n}) \log \frac{\gamma(x_{1..n})}{\mu(x_{1..n})} \le 1 + n\mu(\mathcal{X}^n\setminus T^n_{j_\mu^n}) \log|\mathcal{X}| = o(n), \quad (34)$$

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 (30).

From (32), (33) and (34) 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 (23), 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 (34). ∎

Acknowledgements

Some of the results have appeared (Ryabko, 2010b) in the proceedings of COLT'10. The author is grateful to the anonymous reviewers for their constructive comments on the paper. This research was partially supported by the French Ministry of Higher Education and Research, Nord-Pas-de-Calais Regional Council and FEDER through CPER 2007-2013, ANR projects EXPLO-RA (ANR-08-COSI-004) and Lampada (ANR-09-EMER-007), by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 231495 (project CompLACS), and by Pascal-2.

References

P. Billingsley. *Ergodic theory and information*. Wiley, New York, 1965.

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

N. Cesa-Bianchi and G. Lugosi. On prediction of individual sequences. *Annals of Statistics*, 27:1865–1895, 1999.

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

A. P. Dawid. Prequential data analysis. *Lecture Notes-Monograph Series*, 17:113–126, 1992.

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

R. Krichevsky. *Universal Compression and Retrival*. Kluwer Academic Publishers, 1993.

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

B. Ryabko. Twice-universal coding. *Problems of Information Transmission*, 3:173–177, 1984.

B. Ryabko. Noiseless coding of combinatorial sources, Hausdorff dimension, and Kolmogorov complexity. *Problems of Information Transmission*, 22:16–26, 1986.

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

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.

D. Ryabko. On finding predictors for arbitrary families of processes. *Journal of Machine Learning Research*, 11:581–602, 2010a.

D. Ryabko. Sequence prediction in realizable and non-realizable cases. In *Proc. the 23rd Conference on Learning Theory (COLT 2010)*, pages 119–131, Haifa, Israel, 2010b.

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.

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

A. N. Shiryaev. *Probability*. Springer, 1996.

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

J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24:530–536, 1978.