Sequence prediction in realizable and non-realizable cases

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

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, it is required to give the 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 by total variation distance, then these problems coincide, while if it is measured by 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 $\mathcal{C}$. 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.

1 Introduction

A sequence $x_1, \dots, x_n, \dots$ of discrete-valued observations ($x_i \in \mathcal{X}$, $\mathcal{X}$ is finite) 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, it 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 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 it. 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 $\mathcal{C}$? Therefore, 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 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 [14], the set of k-order Markov and finite-memory processes (e.g. [7]) and the set of all stationary processes [9]. The general question of finding predictors for an arbitrary given set $\mathcal{C}$ of process measures has been addressed in [12, 11]; the latter work shows that when a solution exists it can be obtained as a Bayes mixture over a countable subset of $\mathcal{C}$. There is, however, no algorithm known so far for obtaining a predictor for an arbitrary set $\mathcal{C}$ (when such a predictor it exists).

The non-realizable case is usually studied in a slightly different, non-probabilistic, setting. We refer to [2] for a comprehensive overview. It is assumed that the observed sequence of outcomes is an arbitrary (deterministic) sequence; it is required not to give conditional probabilities, but just guesses of the next outcomes. 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) $\mathcal{C}$. In this approach, it is mostly assumed that the set $\mathcal{C}$ is finite or countable. The case when $\mathcal{C}$ is the set of all i.i.d. process measures

has also been considered, see [3]. 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). In this sense our non-realizable version of the problem is more difficult. Note that even if one assumes the input sequence to be deterministic, optimal predictions may still be probabilistic; in particular, in the interpretation of Kelly [5] the predictor has to make different stakes on different outcomes, in which case if the sequence of outcomes is a priory unknown, he will never put all the capital on one outcome (otherwise he may be ruined). In [10] this approach is taken further to analyze the optimal growth of the rate of increase of capital for a given set of sequences of outcomes, as compared to the performance of computable or finite-automata predictors. At the same time, 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. Aiming at predicting probabilities of outcomes as close as possible to the "correct" ones also allows us to abstract from the actual use of the predictions (e.g. making bets) and thus from considering losses in a general form; instead, we can concentrate on the form of losses (measuring the discrepancy between the forecast and true probabilities) which are more convenient for the analysis. Noteworthy, the probabilistic approach also makes the machinery of probability theory applicable, hopefully making the problem easier.

In this work we consider two measures of the quality of prediction. The first one is 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 [1] and only if [4] 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 analyze. 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 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. The second 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 problem (Problem 3). We then analyze 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 [11] 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. This allows us to obtain analogues of such algebraic properties of the space of process measures ordered with respect to absolute continuity (prediction in total variation) as existence of supremum and infimum of every bounded set (order completeness), in the case of prediction in expected average KL divergence. 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 shown to demonstrate that there is a solution to this problem for the set of all finite-memory process measures. 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 [9] which 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).

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 $\Omega := (\mathcal{X}^\infty, \mathcal{B})$ where $\mathcal{B}$ is the sigma-field generated by the cylinder sets $[x_{1..n}]$, $x_i \in \mathcal{X}, n \in \mathbb{N}$, where $[x_{1..n}]$ is the set of all infinite sequences that start with $x_{1..n}$. For a finite

set A denote $|A|$ its cardinality. We use $\mathbf{E}_\mu$ for expectation with respect to a measure μ.

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

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

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 [1] and only if [4] 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. 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

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

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, [9] 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 [9]) 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.

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

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

$$\mathcal{S} := \{\mu\in\mathcal{P} : \forall n,k\geq 0\, \forall a_{1..n}\in\mathcal{X}^n\, \mu(x_{1..n}=a_{1..n}) = \mu(x_{1+k..n+k}=a_{1..n})\}. \tag{4}$$

$$\mathcal{M}_k := \{\mu\in\mathcal{S} : \forall n\geq 0\, \forall a\in\mathcal{X}\, \mu(x_{n+1}=a|x_{1..n}) = \mu(x_{n+1}=a|x_{n-k+1..n})\}, \tag{5}$$

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

Lemma 1. *For every $\rho\in\mathcal{P}$ there exists $\mu\in\mathcal{D}$ such that $d_n(\mu,\rho)\geq n$ for all $n\in\mathbb{N}$.*

3 Sequence prediction problems

For the two notions of predictive quality introduced, we can now start stating 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 $\mathcal{C}$. The set $\mathcal{C}$ here is a set of measures generating 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.

Recall that a measure ρ predicts μ in total variation if and only if μ is absolutely continuous with respect to ρ. Since the relation $\geq_{tv}$ of absolute continuity is transitive, we immediately get the following statement

Proposition 1. *For the case of prediction in total variation, Problems 1 and 2 coincide: every solution to one of them is also a solution to the other.*

However, the relation "ρ predicts μ in expected average KL divergence" is not transitive, so we cannot make the same statement about it.

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 on ν is going to be, compared to the performance of the predictors from $\mathcal{C}$. To address the fully agnostic case, we have to introduce the notion of loss.

Definition 4. *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: $l(\nu, \rho) \leq l(\nu, \mu)$ 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 [7]). 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 well predicted (in the sense on 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.

The following statement is rather obvious.

Proposition 2. *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

less clear whether the three problems are indeed different. It appears that this depends on the measure of predictive quality chosen.

Theorem 1. *(i) For the case of prediction in total variation distance, Problems 1, 2, and 3 coincide: any solution to any one of them is a solution to the other two.*

(ii) 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.

To prove the first statement we will need the following lemma, which is an easy consequence of [1].

Lemma 2. *Let μ, ρ be two process measures. Then $v(\mu, \rho, x_{1..n})$ converges to either 0 or 1 with μ-probability 1.*

Proof. 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 are such that μ_a is absolutely continuous with respect to ρ and μ_s is singular with respect to ρ. Clearly, μ_a and μ_s are singular with respect to each other; let W be such a set that $\mu_a(W) = \rho(W) = 1$ and $\mu_s(W) = 0$. Assume, w.l.o.g., that $\alpha \in (0,1)$ (the other case is trivial). From [1] 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, $v(\mu_s, \rho, x_{1..n}) \geq |\mu_s(W|x_{1..n}) - \rho_s(W|x_{1..n})| = 1$ so that $v(\mu_s, \rho, x_{1..n}) \to 1$ μ_s-a.s. We have

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

and

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

for $x_{1,\dots} \in W$ we have $I \to 0$ μ-a.s., and for $x_{1,\dots} \notin W$ we have $II \to 1$ μ-a.s., which concludes the proof. □

Proof of Theorem 1. The first statement follows trivially from Lemma 2. To prove the second statement of the theorem, we have to provide two examples. Fix the binary alphabet $\mathcal{X} = \{0,1\}$. For each deterministic sequence $t = t_1, t_2, \dots \in \mathcal{D}$ construct the process measure γ_t as follows: $\gamma_t(x_n = t_n | t_{1..n-1}) := 1 - \frac{1}{n}$ 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 one 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, $d_n(\gamma_t, \gamma) \le 1 = o(n)$. 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 discrete measure there is an element of $\mathcal{C}_1$ which predicts it), while by Lemma 1 for every $\rho \in \mathcal{P}$ there exists $t \in \mathcal{D}$ such that $d_n(t, \rho) \ge n$ 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 $t = t_1, t_2, \dots \in \mathcal{D}$ construct the process measure γ_t as follows: $\gamma'_t(x_n = t_n | t_{1..n-1}) := 2/3$ and for $x_{1..n-1} \ne t_{1..n-1}$ 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 $\mathcal{C}_2 := \{\gamma'_t : t \in \mathcal{X}^\infty\}$. 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 $\limsup \frac{1}{n} d_n(t, \rho) \le \log 3/2 < 1$ which contradicts Lemma 1. □

While the examples provided to prove the second statement of the theorem are artificial, there is at least one 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, see Theorem 7 below.

Using Lemma 2 we could also define *expected* (rather than almost sure) total variation loss of ρ with respect to μ, as the probability that $v(\mu, \rho)$ converges to 1, and reformulate Problem 3 for this notion of loss. However, it is easy to see that for this reformulation (the first statement of) Theorem 1 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 statistical point of view such families are rather large: the assumption that the probabilistic law in question has a density with respect to some (nice) measure is a standard one in statistics. It should also be mentioned that such families can easily be uncountable. On the other hand, even such basic examples as the set of all Bernoulli i.i.d. measures does not allow for a predictor that predicts every measure in total variation. Indeed, all these processes are singular with respect to one another; in particular, each of the non-overlapping sets T_p of all sequences which have limiting fraction p of 0s has probability 1 with respect to one of the measures and 0 with respect to all others; since there are uncountably many of these measures, there is no measure ρ with respect to which they

all would have a density (since such a measure should have $\rho(T_p) > 0$ for all p).

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.

4 Results on Problem 2

In Problem 2 we are concerned with the following relation of dominance: ρ "dominates" μ if ρ predicts every ν such that μ predicts ν. For the case of prediction in total variation, this is just the relation of absolute continuity. For the case of prediction in expected average KL divergence, for now all we can say is that this relation is transitive. Denote it by $\geq^0_{KL}$. Formally, we write $\rho \geq^0_{KL} \mu$ if for every $\nu \in \mathcal{P}$ the equality $\limsup \frac{1}{n} d_n(\nu, \mu) = 0$ implies $\limsup \frac{1}{n} d_n(\rho, \mu) = 0$. In this section we will see which properties of $\geq_{tv}$ hold true for $\geq^0_{KL}$.

Let us first recall some facts we know about $\geq_{tv}$; details can be found, for example, in [8]. 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 $[1/2(\mu_1 + \mu_2)]$. 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. Furthermore, $[\mathcal{P}]_{tv}$ is order complete, that is, every upper-bounded set has an exact upper bound: for every $S \subset [\mathcal{P}]_{tv}$ if there is $\rho \in [\mathcal{P}]_{tv}$ such that $\rho \geq \mu$ for all $\mu \in S$, then there exists $\rho' \in [\mathcal{P}]_{tv}$ such that $\rho' = \sup\{\sigma : \sigma \in S\}$. Moreover, every bounded set of disjoint elements of $[\mathcal{P}]_{tv}$ is countable. (The latter statement gives us a criterion for the existence of a solution to Problems 1-3 for total variation.) This can also be used to derive the following [11]: for every bounded set $S \in [\mathcal{P}]_{tv}$ there is a sequence $\sigma_n \in S$, $n \in \mathbb{N}$, such that $\sum_{n\in\mathbb{N}} \sigma_n = \sup\{\sigma : \sigma \in S\}$; here $\sum_{n\in\mathbb{N}} \sigma_n$ means $[\sum_{n\in\mathbb{N}} w_n s_n]$, where $w_n > 0$ are such that $\sum_{n\in\mathbb{N}} w_n = 1$ and $s_n \in \sigma_n$, $n \in \mathbb{N}$.

The key to establishing a similar theory about $\geq^0_{KL}$ is generalizing the latter fact.

Theorem 2. *Let $\mathcal{C}$ be a set of probability measures on Ω. If there is a measure ρ such that $\rho \geq^0_{KL} \mu$ for every $\mu \in \mathcal{C}$, then there is a sequence $\mu_k \in \mathcal{C}$, $k \in \mathbb{N}$ such that $\sum_{k\in\mathbb{N}} w_k\mu_k \geq^0_{KL} \mu$ for every $\mu \in \mathcal{C}$, where w_k are some positive weights.*

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

$$U^n_\mu := \left\{x_{1..n} \in \mathcal{X}^n : \mu(x_{1..n}) \geq \frac{1}{n}\rho(x_{1..n})\right\}, \tag{6}$$

$$V^n_\mu := \left\{x_{1..n} \in \mathcal{X}^n : p(\mu)(x_{1..n}) \geq 2^{-\delta_n(\mu)}\mu(x_{1..n})\right\}, \tag{7}$$

and

$$T^n_\mu := U^n_\mu \cap V^n_\mu. \tag{8}$$

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

$$\mu(\mathcal{X}^n \backslash U^n_\mu) = \mu\left(\frac{\rho(x_{1..n})}{\mu(x_{1..n})} > n\right) \leq \frac{1}{n} E_\mu \frac{\rho(x_{1..n})}{\mu(x_{1..n})} = \frac{1}{n}. \tag{9}$$

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

$$\begin{aligned} -\sum_{x_{1..n}\in A} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} &= -\mu(A) \sum_{x_{1..n}\in A} \frac{1}{\mu(A)}\mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &\geq -\mu(A) \log \frac{\rho(A)}{\mu(A)} \geq -\mu(A)\log\rho(A) - \frac{1}{2}. \end{aligned} \tag{10}$$

Moreover,

$$d_n(\mu, p(\mu)) = -\sum_{x_{1..n}\in\mathcal{X}^n\setminus V_\mu^n} \mu(x_{1..n}) \log \frac{p(\mu)(x_{1..n})}{\mu(x_{1..n})}$$
$$- \sum_{x_{1..n}\in V_\mu^n} \mu(x_{1..n}) \log \frac{p(\mu)(x_{1..n})}{\mu(x_{1..n})} \geq \delta_n(\mu_n)\mu(\mathcal{X}^n\setminus V_\mu^n) - 1/2,$$

where in the inequality we have used (7) for the first summand and (10) for the second. Thus,

$$\mu(\mathcal{X}^n\setminus V_\mu^n) \leq \frac{d_n(\mu, p(\mu)) + 1/2}{\delta_n(\mu)} = o(1). \tag{11}$$

From (8), (9) and (11) we conclude

$$\mu(\mathcal{X}^n\setminus T_\mu^n) \leq \mu(\mathcal{X}^n\setminus V_\mu^n) + \mu(\mathcal{X}^n\setminus U_\mu^n) = o(1). \tag{12}$$

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 $m_k^n := \max_{\mu\in\mathcal{C}} \rho(T_\mu^n\setminus T_{k-1}^n)$. If $m_k^n > 0$, let μ_k^n be any $\mu \in \mathcal{C}$ such that $\rho(T_{\mu_k^n}^n\setminus T_{k-1}^n) = m_k^n$, 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

$$\nu_n := \sum_{k=1}^{K_n} w_k p(\mu_k^n). \tag{13}$$

As a result of this construction, for every $n \in \mathbb{N}$ every $k \leq K_n$ and every $x_{1..n} \in T_k^n$ using the definitions (8), (6) and (7) we obtain

$$\nu_n(x_{1..n}) \geq w_k \frac{1}{n} 2^{-\delta_n(\mu)} \rho(x_{1..n}). \tag{14}$$

Step 2: the resulting predictor. Finally, define

$$\nu := \frac{1}{2}\gamma + \frac{1}{2}\sum_{n\in\mathbb{N}} w_n \nu_n, \tag{15}$$

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 $\rho(T_k^n \backslash T_{k-1}^n) \geq \rho(T_{k+1}^n \backslash T_k^n)$, 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,

$$\rho(T_\mu^n \backslash T_{j_\mu^n}^n) \leq \varepsilon_\mu^n. \tag{16}$$

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

$$\begin{aligned} d_n(\mu,\rho) = &- \sum_{x_{1..n} \in T_\mu^n \cap T_{j_\mu^n}^n} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &- \sum_{x_{1..n} \in T_\mu^n \backslash T_{j_\mu^n}^n} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &- \sum_{x_{1..n} \in \mathcal{X}^n \backslash T_\mu^n} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \\ &= I + II + III. \end{aligned} \tag{17}$$

Then, from (8) and (6) we get

$$I \geq -\log n. \tag{18}$$

From (10) and (16) we get

$$II \geq -\mu(T_\mu^n \backslash T_{j_\mu^n}^n) \log \rho(T_\mu^n \backslash T_{j_\mu^n}^n) - 1/2 \geq -\mu(T_\mu^n \backslash T_{j_\mu^n}^n) \log \varepsilon_\mu^n - 1/2. \tag{19}$$

Furthermore,

$$\begin{aligned} III \geq \sum_{x_{1..n} \in \mathcal{X}^n \backslash T_\mu^n} \mu(x_{1..n}) \log \mu(x_{1..n}) &\geq \mu(\mathcal{X}^n \backslash T_\mu^n) \log \frac{\mu(\mathcal{X}^n \backslash T_\mu^n)}{|\mathcal{X}^n \backslash T_\mu^n|} \\ &\geq -\frac{1}{2} - \mu(\mathcal{X}^n \backslash T_\mu^n) n \log |\mathcal{X}|, \end{aligned} \tag{20}$$

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 $|\mathcal{X}^n \backslash T_\mu^n| \leq |\mathcal{X}|^n$. Combining (17) with the bounds (18), (19) and (20) we obtain

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

so that

$$\mu(T_\mu^n \backslash T_{j_\mu^n}^n) \leq \frac{1}{-\log \varepsilon_\mu^n} \Big(d_n(\mu, \rho) + \log n + 1 + \mu(\mathcal{X}^n \backslash T_\mu^n) n \log |\mathcal{X}| \Big). \quad (21)$$

From the fact that $d_n(\mu, \rho) = o(n)$ and (12) it follows that the term in brackets is $o(n)$, so that 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 (21) gives $\mu(T_\mu^n \backslash T_{j_\mu^n}^n) = o(1)$. Fix such a choice of ε_μ^n. Then, using (12), we conclude

$$\mu(\mathcal{X}^n \backslash T_{j_\mu^n}^n) \leq \mu(\mathcal{X}^n \backslash T_\mu^n) + \mu(T_\mu^n \backslash T_{j_\mu^n}^n) = o(1). \quad (22)$$

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

$$\begin{aligned} \nu(x_{1..n}) \geq \frac{1}{2} w_n \nu_n(x_{1..n}) &\geq \frac{1}{2} w_n w_{j_\mu^n} \frac{1}{n} 2^{-\delta_n(\mu)} \rho(x_{1..n}) \\ &= \frac{w_n w}{2n} (\varepsilon_\mu^n)^2 2^{-\delta_n(\mu)} \rho(x_{1..n}), \quad (23) \end{aligned}$$

where the first inequality follows from (15), the second from (14), 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

$$\begin{aligned} d_n(\mu, \nu) = &- \sum_{x_{1..n} \in T_{j_\mu^n}^n} \mu(x_{1..n}) \log \frac{\nu(x_{1..n})}{\mu(x_{1..n})} \\ &- \sum_{x_{1..n} \in \mathcal{X}^n \backslash T_{j_\mu^n}^n} \mu(x_{1..n}) \log \frac{\nu(x_{1..n})}{\mu(x_{1..n})} = I + II. \quad (24) \end{aligned}$$

From (23) we find

$$I \le -\log\Big(\frac{w_n w}{2n}(\varepsilon_\mu^n)^2 2^{-\delta_n(\mu)}\Big) - \sum_{x_{1..n}\in T^n_{j^n_\mu}} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})}$$

$$= (1 + 3\log n - 2\log\varepsilon_\mu^n - 2\log w + \delta_n(\mu))$$

$$+ \left(d_n(\mu,\rho) + \sum_{x_{1..n}\in\mathcal{X}^n\setminus T^n_{j^n_\mu}} \mu(x_{1..n}) \log \frac{\rho(x_{1..n})}{\mu(x_{1..n})} \right)$$

$$\le o(n) - \sum_{x_{1..n}\in\mathcal{X}^n\setminus T^n_{j^n_\mu}} \mu(x_{1..n}) \log \mu(x_{1..n})$$

$$\le o(n) + \mu(\mathcal{X}^n\setminus T^n_{j^n_\mu}) n \log|\mathcal{X}| = o(n), \quad (25)$$

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 (22). Moreover, from (15) we find

$$II \le \log 2 - \sum_{x_{1..n}\in\mathcal{X}^n\setminus T^n_{j^n_\mu}} \mu(x_{1..n}) \log \frac{\gamma(x_{1..n})}{\mu(x_{1..n})}$$

$$\le 1 + n\mu(\mathcal{X}^n\setminus T^n_{j^n_\mu}) \log|\mathcal{X}| = o(n), \quad (26)$$

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 (22).

From (24), (25) and (26) 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 (15), 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 (26). $\square$

Denote $[\mathcal{P}]^0_{KL}$ the set of equivalence classes with respect to $\geq^0_{KL}$, and $[\mu]^0_{KL}$ the equivalence class that contains $\mu \in \mathcal{P}$. Let us add a new element 0 to the set $[\mathcal{P}]^0_{KL}$ which by definition satisfies $s \geq^0_{KL} 0$ for all $s \in [\mathcal{P}]^0_{KL}$. From Theorem 2 we can obtain the following corollary.

Theorem 3. *(i) Every pair s_1, s_2 of elements of $[\mathcal{P}]^0_{KL}$ has a supremum $\sup(s_1, s_2)$ and an infimum $\inf(s_1, s_2)$.*

(ii) Every upper-bounded subset of $[\mathcal{P}]^0_{KL}$ has an exact upper bound.

Proof. We start with the second statement. Let $S \subset [\mathcal{P}]^0_{KL}$ be upper-bounded. Then by Theorem 2 there is a sequence $s_k \in \mathcal{S}$, $k \in \mathbb{N}$, such that $\rho := [\sum_{k\in\mathbb{N}} w_k \sigma_k]^0_{KL}$, where w_k are some positive weights and $\sigma_k \in s_k$, is an upper bound for S. Let ρ' be any other upper bound of S. Then clearly $\rho' \geq^0_{KL} s_k$ for all $k \in \mathbb{N}$; but then also $\rho' \geq^0_{KL} \rho$, and the second statement is proven.

To prove the first statement, observe that, for any two elements $[\mu_1]^0_{KL}, [\mu_2]^0_{KL} \in [\mathcal{P}]^0_{KL}$, their supremum is given by $[1/2(\mu_1+\mu_2)]^0_{KL}$.The existence of $\inf(s_1, s_2)$ follows from the second statement of the theorem applied to the set $\{s \in [\mathcal{P}]^0_{KL} : s_1 \geq^0_{KL} s,\ s_2 \geq^0_{KL} s\}$, which is obviously bounded (by s_1 and s_2). $\square$

5 Results on Problem 3

For the third problem we can also introduce a relation on process measures: for $\rho, \mu \in \mathcal{P}$ let $\rho \geq_{KL} \mu$ if $\limsup \frac{1}{n} d_n(\nu, \mu) \geq \limsup \frac{1}{n} d_n(\nu, \rho)$ for every $\nu \in \mathcal{P}$. However, for this relation we do not currently have an analogue of Theorem 2. That is why we take a different route for analysis of Problem 3 for expected average KL divergence.

Again, we start by analogy with prediction in total variation. Knowing that a mixture of a countable subset gives a predictor if there is one, a notion that naturally comes to mind when trying to characterize families of processes for which a predictor exists, is separability. Can we say that there is a solution to Problem 3 for a class $\mathcal{C}$ of process measures if and only if $\mathcal{C}$ is separable? Of course, to talk about separability we need a suitable topology on the space of all measures, or at least on $\mathcal{C}$. If the formulated questions were to have a positive answer, we would need a different topology for each of the notions of predictive quality that we consider. In the case of total variation distance we obviously have a candidate topology: that of total

variation distance, and indeed separability with respect to this topology is equivalent to the existence of a predictor, as the next theorem shows.

Definition 5 (unconditional total variation distance)**.** *Introduce the (unconditional) total variation distance*

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

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

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

Necessity. For any $\mu\in\mathcal{C}$, since ρ predicts μ in total variation, μ has a density (Radon-Nikodym derivative) f_μ with respect to ρ. Thus, for the set $T:=\{x\in X^\infty : \exists\mu\in\mathcal{C}\ f_\mu(x)\neq 0\}$ we have $\mu(T)=1$ for all $\mu\in\mathcal{C}$. We can define L_1 distance with respect to ρ as follows $L_1^\rho(\mu,\nu)=\int_T |f_\mu-f_\nu|d\rho$. The set $\mathcal{C}$ is separable with respect to this distance, for example a dense countable subset $\mathcal{D}$ can be constructed as the set of measures whose densities are step-functions with finitely many steps, that take only rational values (see e.g. [6]). Let $\mathcal{D}$ be any such set. Thus for every $\mu\in\mathcal{C}$ and every ε there is a $\mu'\in\mathcal{D}$ such that $L_1^\rho(\mu,\mu')<\varepsilon$. Then for every measurable set A we have

$$|\mu(A)-\mu'(A)| = |\int_A f_\mu d\rho - \int_A f_{\mu'}d\rho| \leq \int_A |f_\mu-f_{\mu'}|d\rho \leq \int_T |f_\mu-f_{\mu'}|d\rho<\varepsilon.$$

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

In the case of expected average KL divergence the situation is different. While one can introduce a topology based on it, separability with respect to this topology turns out to be a sufficient but not a necessary condition for the existence of a predictor, as is shown in the next theorem.

Definition 6. *Define the distance $d_\infty(\mu_1,\mu_2)$ on process measures as follows*

$$d_\infty(\mu_1,\mu_2)=\limsup_{n\to\infty}\sup_{x_{1..n}\in\mathcal{X}^n}\frac{1}{n}\left|\log\frac{\mu_1(x_{1..n})}{\mu_2(x_{1..n})}\right|. \tag{27}$$

Clearly, d_∞ is symmetric and transitive, but is not exact. Moreover, for every μ_1,μ_2 we have

$$\limsup_{n\to\infty}\frac{1}{n}d_n(\mu_1,\mu_2)\le d_\infty(\mu_1,\mu_2). \tag{28}$$

Theorem 5. *(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 $\nu:=\sum_{k\in\mathbb{N}}w_k\mu_k$. Fix any measure τ and any $\mu\in\mathcal{C}$. We will show that $\limsup_{n\to\infty}\frac{1}{n}d_n(\tau,\nu)\le\limsup_{n\to\infty}\frac{1}{n}d_n(\tau,\nu)$. 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 $\sqrt{n}$. 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 5 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.

Theorem 6. *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 5. It is enough to show that each set $\mathcal{M}_k$ is separable with respect to d_∞.

Observe that the family $\mathcal{M}_k$ of k-order stationary binary-valued Markov processes is parametrized by $|\mathcal{X}|^{k+1}$ $[0,1]$-valued parameters: probability of observing 0 after observing $x_{1..k}$, for each $x_{1..k} \in \mathcal{X}^k$. For each $k \in \mathbb{N}$ let μ^k_q, $q \in Q^{2^k}$ be the (countable) family of all stationary k-order Markov processes with rational values of all the parameters. We will show that this family is dense in $\mathcal{M}_k$. 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

$$\frac{1}{n}\left|\log\frac{\mu_1(x_{1..n})}{\mu_2(x_{1..n})}\right| \leq 2\log(a+\tau) \tag{29}$$

where $a = \inf_{x_{1..k}:\mu_i(x_{1..k})\neq 0, i=1,2}\mu_i(x_{1..k})$ and $\tau := \inf_{x\in\mathcal{X},x_{1..k}\in\mathcal{X}^k}|\mu_1(x|x_{1..k} - \mu_2(x|x_{1..k})|$. Since the set μ^k_q, $q \in Q^{2^k}$ is dense in $\mathcal{M}_k$ with respect to this parametrization, the for each $\mu \in M_k$ the expression (29) can be made arbitrary small for appropriate μ^k_q, so that $\mathcal{M}_k$ is separable with respect to d_∞. □

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 we have the following.

Theorem 7. *There is [9] a solution to Problem 1 for the set* $\mathcal{S}$ *of all stationary process measures, for the case of prediction in expected average KL divergence. There is no solution to Problem 3 for* $\mathcal{S}$.

Proof. The following proof of the second statement is based on the construction similar to the one used in [9] 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. [13]); taken as the initial distribution it defines a stationary ergodic process with values in $\mathbb{N} \cup \{0,1\}$. 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 the sequence t.

To show that there is no solution to Problem 3 for $\mathcal{S}$, we will show that there is no solution for 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 1. □

References

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

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

[3] Y. Freund. Predicting a binary sequence almost as well as the optimal biased coin. *Information and Computation*, 182(2):73–94, 2003.

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

[5] J.L. Kelly. A new interpretation of information rate. *IRE Transactions on Information Theory*, 2:185–189, 1956.

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

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

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

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

[10] B. Ryabko. The complexity and effectiveness of prediction algorithms. *Journal of Complexity*, 10:281–295, 1994.

[11] 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.

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

[13] A. N. Shiryaev. *Probability*. Springer, 1996.

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