Sequence prediction for non-stationary processes

Daniil Ryabko and **Marcus Hutter**

IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland*
{daniil,marcus}@idsia.ch, http://www.idsia.ch/~{daniil,marcus}

Abstract

We address the problem of sequence prediction for nonstationary stochastic processes. In particular, given two measures on the set of one-way infinite sequences over a finite alphabet, consider the question whether one of the measures predicts the other. We find some conditions on local absolute continuity under which prediction is possible.

1 Introduction

Let a sequence x_t, $t \in I\!N$ of letters from some finite alphabet $\mathcal{X}$ be generated by some probability measure μ. Having observed the first n letters $x_1,...,x_n$ we want to predict what is the probability of the next letter being x, for each $x \in \mathcal{X}$. This task is motivated by numerous applications — from weather forecasting and stock market prediction to data compression. It also generalizes to the problem of reinforcement learning in an arbitrary (non-Markov) environment. Indeed, in the sequence prediction problem Markov measures generalize to stationary and ergodic measures, for which prediction is possible, as will be explained below. However, such generalization is not possible in active (reinforcement) learning problems, and thus one has to look for probabilistic conditions that generalize.

If the measure μ is known completely then the best forecasts one can make for the $(n+1)$st outcome of a sequence $x_1,...,x_n$ is μ-conditional probabilities of x_{n+1} being $x \in \mathcal{X}$ given $x_1,...,x_n$. On the other hand, it is immediately apparent that if nothing is known about the distribution μ generating the sequence then no prediction is possible, since for any predictor there is a measure on which it errs (gives grossly wrong probability forecasts) on every step. Thus one has to restrict the attention to some class of measures. Laplace was perhaps the first to address the question of sequence prediction, his motivation being as follows: Suppose that we know that the Sun has risen every day for 5000 years, what is the probability that it will rise tomorrow? He suggested to assume that the probability that the Sun rises is the same every day and the trials are

*This work was supported by the Swiss NSF grant 200020-107616.

independent of each other. Thus Laplace considered the task of sequence prediction when the true generating measure belongs to the family of Bernoulli i.i.d. measures with binary alphabet $\mathcal{X}=\{0,1\}$. The predicting measure suggested by Laplace was $\rho_L(x_{n+1}=1|x_1,...,x_n)=\frac{k+1}{n+2}$ where k is the number of 1s in $x_1,...,x_n$. The conditional probabilities of Laplace's measure ρ_L converge to the true conditional probabilities μ-almost surely under any Bernoulli i.i.d measure μ. This approach generalizes to the problem of predicting any finite-memory (e.g. Markovian) measure. Moreover, in [4] a measure ρ_R was constructed for predicting an arbitrary stationary measure. The conditional probabilities of ρ_R converge to the true ones *on average*, where the average is taken over time steps (that is, in Cesaro sense), μ-almost surely for any stationary measure μ. However, as it was shown in the same work, there is no measure for which conditional probabilities converge to the true ones μ-a.s. for every stationary μ. Thus we can see that already for the problem of predicting outcomes of a stationary measure two criteria of prediction arise: prediction in the average (or in Cesaro sense) and prediction on each step, and the solution exists only for the former problem.

But what if the measure generating the sequence is not stationary? A different assumption one can make is that the measure μ generating the sequence is computable. Solomonoff [6, Eq.(13)] suggested a measure ξ for predicting any computable probability measure. The key observation here is that the class of all computable probability measures is countable; let us denote it by $(\nu_i)_{i\in\mathbb{N}}$. A Bayesian predictor ξ for a countable class of measures $(\nu_i)_{i\in\mathbb{N}}$ is constructed as follows: $\xi(A)=\sum_{i=1}^{\infty}w_i\nu_i(A)$ for any measurable set A, where the weights w_i are positive and sum to one[1]. The best predictor for a measure μ is the measure μ itself. The Bayesian predictor simply takes the weighted average of the predictors for all measures in the class — for countable classes this is possible. It was shown by Solomonoff [7] that ξ-conditional probabilities converge to μ-conditional probabilities almost surely for any computable measure μ. In fact this is a special case of a more general (though without convergence rate) result of Blackwell and Dubins [1] which states that if a measure μ is absolutely continuous with respect to a measure ρ then the conditional measure ρ given $x_1,...,x_n$ converges to μ given $x_1,...,x_n$ in total variation μ-almost surely. Convergence in total variation means prediction in a very strong sense — convergence of probabilities of arbitrary events (not just the next outcome), or prediction with arbitrary fast growing horizon. Since for ξ we have $\xi(A)\geq w_i\nu_i(A)$ for every measurable set A and for every ν_i, each ν_i is absolutely continuous with respect to ξ.

Thus the problem of sequence prediction for certain classes of measures (such as the class of all stationary measures or the class of all computable measures) was often addressed in the literature. Although the mentioned classes of measures are sufficiently interesting, it is often hard to decide in applications with

[1] It is not necessary for prediction that the weights sum to one. In [7] and [8] $w_i=2^{-K(i)}$ where K stands for the prefix Kolmogorov complexity, and so the weights do not sum to 1. Further, the ν and ξ are only semi-measures.

which assumptions does a problem at hand comply; not to mention such practical issues as that a predicting measure for all computable measures is necessarily non-computable itself. Moreover, to be able to generalize the solutions of the sequence prediction problem to such problems as active learning, where outcomes of a sequence may depend on actions of the predictor, one has to understand better under which conditions the problem of sequence prediction is solvable. In particular, in active learning, the stationarity assumption does not seem to be applicable (since the predictions are non-stationary), although, say, the Markov assumption is often applicable and is extensively studied. Thus, we formulate the following general questions which we start to address in the present work:

General motivating questions. For which classes of measures is sequence prediction possible? Under which conditions does a measure ρ predict a measure μ?

As we have seen, these questions have many facets, and in particular there are many criteria of prediction to be considered, such as almost sure convergence of conditional probabilities, convergence in average, etc. Extensive as the literature on sequence prediction is, these questions in their full generality have not received much attention. One line of research which exhibits this kind of generality consists in extending the result of Blackwell and Dubins mentioned above, which states that if μ is absolutely continuous with respect to ρ, then ρ predicts μ in total variation distance. In [3] a question of whether, given a class of measures $\mathcal{C}$ and a prior ("meta"-measure) λ over this class of measures, the conditional probabilities of a Bayesian mixture of the class $\mathcal{C}$ w.r.t. λ converge to the true μ-probabilities (weakly merge, in terminology of [3]) for λ-almost any measure μ in $\mathcal{C}$. This question can be considered solved, since the authors provide necessary and sufficient conditions on the measure given by the mixture of the class $\mathcal{C}$ w.r.t. λ under which prediction is possible. The major difference from the general questions we posed above is that we do not wish to assume that we have a measure on our class of measures. For large (non-parametric) classes of measures it may not be intuitive which measure over it is natural; rather, the question is whether a "natural" measure which can be used for prediction exists.

To address the general questions posed, we start with the following observation. As it was mentioned, for a Bayesian mixture ξ of a countable class of measures ν_i, $i\in I\!N$, we have $\xi(A)\geq w_i\nu_i(A)$ for any i and any measurable set A, where w_i is a constant. This condition is stronger than the assumption of absolute continuity and is sufficient for prediction in a very strong sense. Since we are willing to be satisfied with prediction in a weaker sense (e.g. convergence of conditional probabilities), let us make a weaker assumption: Say that *a measure ρ dominates a measure μ with coefficients $c_n>0$* if

$$\rho(x_1,\dots,x_n) \;\geq\; c_n\mu(x_1,\dots,x_n) \qquad (1)$$

for all x_1,...,x_n.

The concrete question we pose is, under what conditions on c_n does (1) imply that ρ predicts μ? Observe that if $\rho(x_1,...,x_n)>0$ for any x_1,...,x_n then any measure μ is *locally* absolutely continuous with respect to ρ (that is, the measure μ restricted to the first n trials $\mu|_{\mathcal{X}^n}$ is absolutely continuous w.r.t. $\rho|_{\mathcal{X}^n}$ for each n), and moreover, for any measure μ some constants c_n can be found that satisfy (1). For example, if ρ is Bernoulli i.i.d. measure with parameter $\frac{1}{2}$ and μ is any other measure, then (1) is (trivially) satisfied with $c_n = 2^{-n}$. Thus we know that if $c_n \equiv c$ then ρ predicts μ in a very strong sense, whereas exponentially decreasing c_n are not enough for prediction. Perhaps somewhat surprisingly, we will show that dominance with any subexponentially decreasing coefficients is sufficient for prediction, in a weak sense of convergence of expected averages. Dominance with any polynomially decreasing coefficients, and also with coefficients decreasing (for example) as $c_n = \exp(-\sqrt{n}/\log n)$, is sufficient for (almost sure) prediction on average (i.e. in Cesaro sense). However, for prediction on every step we have a negative result: for any dominance coefficients that go to zero there exists a pair of measures ρ and μ which satisfy (1) but ρ does not predict μ in the sense of almost sure convergence of probabilities. Thus the situation is similar to that for predicting any stationary measure: prediction is possible in the average but not on every step.

Note also that for Laplace's measure ρ_L it can be shown that ρ_L dominates any i.i.d. measure μ with linearly decreasing coefficients $c_n = \frac{1}{n+1}$; a generalization of ρ_L for predicting all measures with memory k (for a given k) dominates them with polynomially decreasing coefficients. Thus dominance with decreasing coefficients generalizes (in a sense) predicting countable classes of measures (where we have dominance with a constant), absolute continuity (via local absolute continuity), and predicting i.i.d. and finite-memory measures.

2 Notation and Definitions

We consider processes on the set of one-way infinite sequences $\mathcal{X}^\infty$ where $\mathcal{X}$ is a finite set (alphabet). In the examples we will often assume $\mathcal{X} = \{0,1\}$. The notation $x_{1:n}$ is used for $x_1,\dots,x_n$ and $x_{<n}$ for $x_1,\dots,x_{n-1}$, $x_t \in \mathcal{X}$. The symbol μ is reserved for the "true" measure generating examples. We use $\mathbf{E}_\nu$ for expectation with respect to a measure ν and simply $\mathbf{E}$ for $\mathbf{E}_\mu$ (expectation with respect to the "true" measure generating examples).

For two measures μ and ρ define the following measures of divergence.

(d) Kullblack-Leibler (KL) divergence

$$d_n(\mu,\rho|x_{<n}) = \sum_{x\in\mathcal{X}} \mu(x_n = x|x_{<n}) \log \frac{\mu(x_n = x|x_{<n})}{\rho(x_n = x|x_{<n})},$$

($\bar{d}$) average KL divergence

$$\bar{d}_n(\mu,\rho|x_{1:n}) = \frac{1}{n}\sum_{t=1}^{n} d_t(\mu,\rho|x_{<n}),$$

(a) absolute distance

$$a_n(\mu,\rho|x_{<n}) = \sum_{x\in\mathcal{X}} |\mu(x_n = x|x_{<n}) - \rho(x_n = x|x_{<n})|,$$

$(\bar{a})$ average absolute distance

$$\bar{a}_n(\mu,\rho|x_{1:n}) = \frac{1}{n}\sum_{t=1}^{n} a_t(\mu,\rho|x_{<n}).$$

Definition 1 (Convergence concepts) *We say that ρ predicts μ*

(d) *in KL divergence if $d_n(\mu,\rho|x_{<n}) \to 0$ μ-a.s. as $t\to\infty$,*

$(\bar{d})$ *in average KL divergence if $\bar{d}_n(\mu,\rho|x_{1:n}) \to 0$ μ-a.s.,*

$(\mathbf{E}\bar{d})$ *in expected average KL divergence if $\mathbf{E}_\mu \bar{d}_n(\mu,\rho|x_{1:n}) \to 0$,*

(a) *in absolute distance if $a_n(\mu,\rho|x_{<n}) \to 0$ μ-a.s.,*

$(\bar{a})$ *in average absolute distance if $\bar{a}_n(\mu,\rho|x_{1:n}) \to 0$ μ-a.s.,*

$(\mathbf{E}\bar{a})$ *in expected average absolute distance if $\mathbf{E}_\mu \bar{a}_n(\mu,\rho|x_{1:n}) \to 0$.*

The argument $x_{1:n}$ will be often left implicit in our notation. A measure ρ converges to a measure μ in *total variation* (tv) if $\sup_{A\subset\sigma(\bigcup_{t=n}^{\infty}\mathcal{X}^t)} |\mu(A|x_{<n}) - \rho(A|x_{<n})| \to 0$ μ-almost surely. Some other measures of prediction ability are considered in Section 4. The following implications hold (and are complete and strict):

$$\begin{tikzcd}
 & d \arrow[r, Rightarrow] \arrow[d, Rightarrow] & \bar{d} \arrow[d, Rightarrow] & \mathbf{E}\bar{d} \arrow[d, Rightarrow] \\
tv \arrow[r, Rightarrow] & a \arrow[r, Rightarrow] & \bar{a} \arrow[r, Rightarrow] & \mathbf{E}\bar{a}
\end{tikzcd}$$

to be understood as e.g.: if $\bar{d}_n \to 0$ a.s. then $\bar{a}_n \to 0$ a.s, or, if $\mathbf{E}\bar{d}_n \to 0$ then $\mathbf{E}\bar{a}_n \to 0$. The horizontal implications $\Rightarrow$ follow immediately from the definitions, and the $\Downarrow$ follow from the following Lemma:

Lemma 2 ($\boldsymbol{a^2 \leq 2d}$) *For all measures ρ and μ and sequences $x_{1:\infty}$ we have: $a_t^2 \leq 2d_t$ and $\bar{a}_n^2 \leq 2\bar{d}_n$ and $(\mathbf{E}\bar{a}_n)^2 \leq 2\mathbf{E}\bar{d}_n$.*

Proof. Pinsker's inequality [2, Lem.3.11*a*] implies $a_t^2 \leq 2d_t$. Using this and Jensen's inequality for the average $\frac{1}{n}\sum_{t=1}^{n}[...]$ we get

$$2\bar{d}_n \;=\; \frac{1}{n}\sum_{t=1}^{n} 2d_t \;\geq\; \frac{1}{n}\sum_{t=1}^{n} a_t^2 \;\geq\; \left(\frac{1}{n}\sum_{t=1}^{n} a_t\right)^2 \;=\; \bar{a}_n^2$$

Using this and Jensen's inequality for the expectation $\mathbf{E}$ we get $2\mathbf{E}\bar{d}_n \geq \mathbf{E}\bar{a}_n^2 \geq (\mathbf{E}\bar{a}_n)^2$. ■

3 Main results

First we consider the question whether property (1) is sufficient for prediction.

Definition 3 (Dominance) *We say that a measure ρ dominates a measure μ with coefficients $c_n > 0$ iff*

$$\rho(x_{1:n}) \;\geq\; c_n \mu(x_{1:n}).$$

for all $x_{1:n}$.

Suppose that ρ dominates μ with decreasing coefficients c_n. Does ρ predict μ in (expected, expected average) KL divergence (absolute distance)? First let us give an example.

Proposition 4 (Dominance of Laplace's measure) *Let ρ_L be the Laplace measure, given by $\rho_L(x_{n+1} = a | x_{1:n}) = \frac{k+1}{n+|\mathcal{X}|}$ for any $a \in \mathcal{X}$ and any $x_{1:n} \in \mathcal{X}^n$, where k is the number of occurrences of a in $x_{1:n}$ (this is also well defined for $n=0$). Then*

$$\rho_L(x_{1:n}) \;\geq\; \frac{n!}{(n+|\mathcal{X}|-1)!}\,\mu(x_{1:n})$$

for any measure μ which generates independently and identically distributed symbols. This bound is sharp.

Proof. We will only give the proof for $\mathcal{X} = \{0,1\}$, the general case is analogous. To calculate $\rho_L(x_{1:n})$ observe that it only depends on the number of 0s and 1s in $x_{1:n}$ and not on their order. Thus we compute $\rho_L(x_{1:n}) = \frac{k!(n-k)!}{(n+1)!}$ where k is the number of 1s. For any measure μ such that $\mu(x_n = 1) = p$ for some $p \in [0,1]$ independently for all n, and for Laplace measure ρ_L we have

$$\begin{aligned} \frac{\mu(x_{1:n})}{\rho_L(x_{1:n})} \;&=\; \frac{(n+1)!}{k!(n-k)!} p^k (1-p)^{n-k} \\ &=\; (n+1)\binom{n}{k} p^k (1-p)^{n-k} \\ &\leq\; (n+1)\sum_{k=0}^{n}\binom{n}{k} p^k (1-p)^{n-k} \;=\; n+1, \end{aligned}$$

for any n-letter word $x_1,...,x_n$ where k is the number of 1s in it. The bound is attained when $p=1$, so that $k=n$, $\mu(x_{1:n})=1$, and $\rho_L(x_{1:n}) = \frac{1}{n+1}$. ■

Thus for Laplace's measure ρ_L and binary $\mathcal{X}$ we have $c_n = \mathcal{O}(\frac{1}{n})$. As mentioned in the introduction, in general, exponentially decreasing coefficients c_n are not sufficient for prediction, since (1) is satisfied with ρ being a Bernoulli i.i.d. measure and μ any other measure. On the other hand, the following proposition shows that in a weak sense of convergence in expected average KL divergence (or absolute distance) the property (1) with subexponentially decreasing c_n is sufficient. We also remind that if c_n are bounded from below then prediction in the strong sense of total variation is possible.

Theorem 5 ($\mathbf{E}\bar{\boldsymbol{d}}\to\mathbf{0}$ and $\mathbf{E}\bar{\boldsymbol{a}}\to\mathbf{0}$) *Let μ and ρ be two measures on $\mathcal{X}^\infty$ and suppose that $\rho(x_{1:n})\geq c_n\mu(x_{1:n})$ for any $x_{1:n}$, where c_n are positive constants satisfying $\frac{1}{n}\log c_n^{-1}\to 0$. Then ρ predicts μ in expected average KL divergence $\mathbf{E}_\mu\bar{d}_n(\mu,\rho)\to 0$ and in expected average absolute distance $\mathbf{E}_\mu\bar{a}_n(\mu,\rho)\to 0$.*

The proof of this theorem is based on the same idea as the proof of convergence of Solomonoff predictor to any of its summands in [4], see also [2].

Proof. For convergence in average expected KL divergence we have

$$\begin{aligned}\mathbf{E}_\mu\bar{d}_n(\mu,\rho)&=\frac{1}{n}\mathbf{E}\sum_{t=1}^n\sum_{x_t\in\mathcal{X}}\mu(x_t|x_{<t})\log\frac{\mu(x_t|x_{<t})}{\rho(x_t|x_{<t})}\\&=\frac{1}{n}\sum_{t=1}^n\mathbf{E}\mathbf{E}^t\log\frac{\mu(x_t|x_{<t})}{\rho(x_t|x_{<t})}=\frac{1}{n}\mathbf{E}\log\prod_{t=1}^n\frac{\mu(x_t|x_{<t})}{\rho(x_t|x_{<t})}\\&=\frac{1}{n}\mathbf{E}\log\frac{\mu(x_{1:n})}{\rho(x_{1:n})}\leq\frac{1}{n}\log c_n^{-1}\to 0,\end{aligned}$$

where $\mathbf{E}^t$ stands for the μ-expectation over x_t conditional on $x_{<t}$.

The statement for expected average distance follows from this and Lemma 2. ■

With a stronger condition on c_n prediction in average KL divergence can be established.

Theorem 6 ($\bar{\boldsymbol{d}}\to\mathbf{0}$ and $\bar{\boldsymbol{a}}\to\mathbf{0}$) *Let μ and ρ be two measures on $\mathcal{X}^\infty$ and suppose that $\rho(x_{1:n})\geq c_n\mu(x_{1:n})$ for every $x_{1:n}$, where c_n are positive constants satisfying*

$$\sum_{n=1}^\infty\frac{(\log c_n^{-1})^2}{n^2}\;<\;\infty. \tag{2}$$

Then ρ predicts μ in average KL divergence $\bar{d}_n(\mu,\rho)\to 0$ μ-a.s. and in average absolute distance $\bar{a}_n(\mu,\rho)\to 0$ μ-a.s.

In particular, the condition (2) on the coefficients is satisfied for polynomially decreasing coefficients, or for $c_n=\exp(-\sqrt{n}/\log n)$.

Proof. Again the second statement (about absolute distance) follows from the first one and Lemma 2, so that we only have to prove the statement about KL divergence.

Introduce the symbol $\mathbf{E}^n$ for μ-expectation over x_n conditional on $x_{<n}$. Consider random variables $l_n=\log\frac{\mu(x_n|x_{<n})}{\rho(x_n|x_{<n})}$ and $\bar{l}_n=\frac{1}{n}\sum_{t=1}^n l_t$. Observe that $d_n=\mathbf{E}^n l_n$, so that the random variables $m_n=l_n-d_n$ form a martingale difference sequence (that is, $\mathbf{E}^n m_n=0$) with respect to the standard filtration defined by $x_1,...,x_n,...$. Let also $\bar{m}_n=\frac{1}{n}\sum_{t=1}^n m_t$. We will show that $\bar{m}_n\to 0$ μ-a.s. and $\bar{l}_n\to 0$ μ-a.s. which implies $\bar{d}_n\to 0$ μ-a.s.

Note that

$$\bar{l}_n\;=\;\frac{1}{n}\log\frac{\mu(x_{1:n})}{\rho(x_{1:n})}\;\leq\;\frac{\log c_n^{-1}}{n}\;\to\;0.$$

Thus to show that $\bar{l}_n$ goes to 0 we need to bound it from below. It is easy to see that $n\bar{l}_n$ is (μ-a.s.) bounded from below by a constant, since $\frac{\rho(x_{1:n})}{\mu(x_{1:n})}$ is a positive μ-martingale whose expectation is 1, and so it converges to a finite limit μ-a.s. by Doob's submartingale convergence theorem, see e.g. [5, p.508].

Next we will show that $\bar{m}_n \to 0$ μ-a.s. We have

$$m_n = \log\frac{\mu(x_{1:n})}{\rho(x_{1:n})} - \log\frac{\mu(x_{<n})}{\rho(x_{<n})} \\ -\mathbf{E}^n \log\frac{\mu(x_{1:n})}{\rho(x_{1:n})} + \mathbf{E}^n \log\frac{\mu(x_{<n})}{\rho(x_{<n})} \\ = \log\frac{\mu(x_{1:n})}{\rho(x_{1:n})} - \mathbf{E}^n \log\frac{\mu(x_{1:n})}{\rho(x_{1:n})}.$$

Let $f(n)$ be some function monotonically increasing to infinity such that

$$\sum_{n=1}^{\infty} \frac{(\log c_n^{-1} + f(n))^2}{n^2} \;<\; \infty \tag{3}$$

(e.g. choose $f(n) = \log n$ and exploit $(\log c_n^{-1} + f(n))^2 \le 2(\log c_n^{-1})^2 + 2f(n)^2$ and (2).) For a sequence of random variables λ_n define

$$(\lambda_n)^{+(f)} \;=\; \begin{cases} \lambda_n & \text{if } \lambda_n \ge -f(n) \\ 0 & \text{otherwise} \end{cases}$$

and $\lambda_n^{-(f)} = \lambda_n - \lambda_n^{+(f)}$. Introduce also

$$m_n^+ \;=\; \left(\log\frac{\mu(x_{1:n})}{\rho(x_{1:n})}\right)^{+(f)} - \mathbf{E}^n \left(\log\frac{\mu(x_{1:n})}{\rho(x_{1:n})}\right)^{+(f)},$$

$m_n^- = m_n - m_n^+$ and the averages $\bar{m}_n^+$ and $\bar{m}_n^-$. Observe that m_n^+ is a martingale difference sequence. Hence to establish the convergence $\bar{m}_n^+ \to 0$ we can use the martingale strong law of large numbers [5, p.501], which states that, for a martingale difference sequence γ_n, if $\mathbf{E}(n\bar{\gamma}_n)^2 < \infty$ and $\sum_{n=1}^{\infty} \mathbf{E}\gamma_n^2/n^2 < \infty$ then $\bar{\gamma}_n \to 0$ a.s. Indeed, for m_n^+ the first condition is trivially satisfied (since the expectation in question is a finite sum of finite numbers), and the second follows from the fact that $|m_n^+| \le \log c_n^{-1} + f(n)$ and (3).

Furthermore, we have

$$m_n^- \;=\; \left(\log\frac{\mu(x_{1:n})}{\rho(x_{1:n})}\right)^{-(f)} - \mathbf{E}^n \left(\log\frac{\mu(x_{1:n})}{\rho(x_{1:n})}\right)^{-(f)}.$$

As it was mentioned before, $\log\frac{\mu(x_{1:n})}{\rho(x_{1:n})}$ converges μ-a.s. either to (positive) infinity or to a finite number. Hence $\left(\log\frac{\mu(x_{1:n})}{\rho(x_{1:n})}\right)^{-(f)}$ is non-zero only a finite number of times, and so its average goes to zero. To see that $\mathbf{E}^n \left(\log\frac{\mu(x_{1:n})}{\rho(x_{1:n})}\right)^{-(f)} \to 0$ we write

$$\begin{aligned}
&\mathbf{E}^n\left(\log\frac{\mu(x_{1:n})}{\rho(x_{1:n})}\right)^{-(f)}\\
&\qquad =\sum_{x_n\in\mathcal{X}}\mu(x_n|x_{<n})\left(\log\frac{\mu(x_{<n})}{\rho(x_{<n})}+\log\frac{\mu(x_n|x_{<n})}{\rho(x_n|x_{<n})}\right)^{-(f)}\\
&\qquad\qquad \geq \sum_{x_n\in\mathcal{X}}\mu(x_n|x_{<n})\left(\log\frac{\mu(x_{<n})}{\rho(x_{<n})}+\log\mu(x_n|x_{<n})\right)^{-(f)}
\end{aligned}$$

and note that the first term in brackets is bounded from below, and so for the sum in brackets to be less than $-f(n)$ (which is unbounded) the second term $\log\mu(x_n|x_{<n})$ has to go to $-\infty$, but then the expectation goes to zero since $\lim_{u\to 0}u\log u=0$.

Thus we conclude that $\bar{m}_n^-\to 0$ μ-a.s., which together with $\bar{m}_n^+\to 0$ μ-a.s. implies $\bar{m}_n\to 0$ μ-a.s., which, finally, together with $\bar{l}_n\to 0$ μ-a.s. implies $\bar{d}_n\to 0$ μ-a.s. ■

However, no form of dominance with decreasing coefficients is sufficient for prediction in absolute distance or KL divergence, as the following negative result states.

Proposition 7 ($d\not\to 0$ and $a\not\to 0$) *For each sequence of positive numbers c_n that goes to 0 there exist measures μ and ρ and a number $\epsilon>0$ such that $\rho(x_{1:n})\geq c_n\mu(x_{1:n})$ for all $x_{1:n}$, yet $a_n(\mu,\rho|x_{1:n})>\epsilon$ and $d_n(\mu,\rho|x_{1:n})>\epsilon$ infinitely often μ-a.s.*

Proof. Let μ be concentrated on the sequence 11111... (that is $\mu(x_n=1)=1$ for all n), and let $\rho(x_n=1)=1$ for all n except for a subsequence of steps $n=n_k$, $k\in I\!N$ on which $\rho(x_{n_k}=1)=1/2$ independently of each other. It is easy to see that choosing n_k sparse enough we can make $\rho(1_1...1_n)$ decrease to 0 arbitrary slowly; yet $|\mu(x_{n_k})-\rho(x_{n_k})|=1/2$ for all k. ■

Thus for the first question — whether dominance with some coefficients decreasing to zero is sufficient for prediction, we have the following table of questions and answers, where, in fact, positive answers for a_n are implied by positive answers for d_n and vice versa for the negative answers:

$\mathbf{E}\bar{d}_n$	$\bar{d}_n$	d_n	$\mathbf{E}\bar{a}_n$	$\bar{a}_n$	a_n
+	+	−	+	+	−

However, if we take into account the conditions on the coefficients, we see some open problems left, and different answers for $\bar{d}_n$ and $\bar{a}_n$ may be obtained. Following is the table of conditions on dominance coefficients and answers to the questions whether these conditions are sufficient for prediction (coefficients bounded from below are included for the sake of completeness).

	$\mathbf{E}\bar{d}_n$	$\bar{d}_n$	d_n	$\mathbf{E}\bar{a}_n$	$\bar{a}_n$	a_n
$\log c_n^{-1}=o(n)$	$+$	?	$-$	$+$	?	$-$
$\sum_{n=1}^{\infty}\frac{\log c_n^{-1}}{n^2}<\infty$	$+$	$+$	$-$	$+$	$+$	$-$
$c_n\geq c>0$	$+$	$+$	$+$	$+$	$+$	$+$

We know from Proposition 7 that the condition $c_n\geq c>0$ for convergence in d_n can not be improved; thus the open problem left is to find whether $\log c_n^{-1}=o(n)$ is sufficient for prediction in $\bar{d}_n$ or at least in $\bar{a}_n$. We conjecture that the answer to the first question is negative and to the second it is positive.

Conjecture 8 ($\bar{\boldsymbol{d}}\not\to\mathbf{0}$) *i) There exist a sequence of numbers c_n that (monotonically) goes to 0, measures μ and ρ and a number $\epsilon>0$ such that $\log c_n^{-1}=o(n)$, $\rho(x_{1:n})\geq c_n\mu(x_{1:n})$ for any $x_{1:n}$, yet $\bar{d}_n(\mu,\rho|x_{1:n})>\epsilon$ infinitely often μ-a.s.*

ii) Suppose a measure ρ dominates a measure μ with such coefficients c_n that $\log c_n^{-1}=o(n)$. Then ρ predicts μ in average absolute distance $\bar{a}_n(\mu,\rho)\to 0$ μ-a.s.

Another open problem is to find out whether any conditions on dominance coefficients are necessary for prediction; so far we only have some sufficient conditions. On the one hand, the obtained results suggest that some form of dominance with decreasing coefficients may be necessary for prediction, at least in the sense of convergence of averages. On the other hand, the condition (1) is uniform over all sequences which probably is not necessary for prediction. As for prediction in the sense of almost sure convergence, perhaps more subtle behavior of the ratio $\frac{\mu(x_{1:n})}{\rho(x_{1:n})}$ should be analyzed, since dominance with decreasing coefficients is not sufficient for prediction in this sense.

4 Miscellaneous

Special cases. In Section 3 we have shown that Laplace's measure ρ_L for $\mathcal{X}=\{0,1\}$ dominates any Bernoulli i.i.d. measure with linearly decreasing coefficients. It can also be shown that a generalization of ρ_L to a measure ρ_L^k for predicting any measure with memory k, for a given k, dominates any such measure with polynomially decreasing coefficients (namely, $c_n^{-1}=\mathcal{O}(n^{|\mathcal{X}|^k})$. The measure ρ_R from [4] for predicting any stationary measure was constructed as a sum of ρ_L^k with positive weights: $\rho_R(x_{1..n})=\sum_{k=1}^{\infty}w_k\rho_L^k(x_{1..n})$. By construction, ρ_R dominates any finite memory measure with polynomially decreasing coefficients. It is interesting to find whether ρ_R (or any other measure which predicts all stationary measures) dominates every stationary measure with some subexponentially decreasing coefficients (or at least dominates non-uniformly). Clearly, this is a special case of the general open question — whether some form of dominance with decreasing coefficients is necessary for prediction.

Other measures of divergence. The last question we discuss is criteria of prediction other than introduced in Section 2. Apart form the measures of divergence of probability measures that we considered we mention also the following:

(s) squared distance

$$s_n(\mu,\rho|x_{<n}) = \sum_{x\in\mathcal{X}} (\mu(x_n = x|x_{<n}) - \rho(x_n = x|x_{<n})^2,$$

(h) Hellinger distance

$$h_n(\mu,\rho|x_{<n}) = \sum_{x\in\mathcal{X}} (\sqrt{\mu(x_n = x|x_{<n})} - \sqrt{\rho(x_n = x|x_{<n}})^2;$$

the average squared distance $\bar{s}_n$ and the average Hellinger distance $\bar{h}_n$ are introduced analogously to $\bar{a}_n$ and $\bar{d}_n$. It is easy to check that all negative results obtained hold with respect to s_n and h_n as well. Positive results for s_n and h_n follow from corresponding positive results for KL divergence d_n and inequalities $s_n(\mu,\rho) \leq d_n(\mu,\rho)$ and $h_n(\mu,\rho) \leq d_n(\mu,\rho)$, see e.g. [2, Lem.3.11]. Expected absolute convergence $\mathbf{E}a_n \to 0$ (also called convergence in the mean) and expected KL convergence $\mathbf{E}d_n \to 0$ may also be considered.

5 Outlook and Conclusion

In the present work we formulated and started to address the question for which classes of measures sequence prediction is possible. Towards this aim we defined the notion of dominance with decreasing coefficients (a condition on local absolute continuity) and found some forms of it which are sufficient for prediction. Besides the more concrete open problems posed, a general program for answering the general questions formulated can be outlined as follows: We would like to find some conditions on dominance with decreasing coefficients which are necessary and sufficient for prediction; for those notions of prediction ability for which this is not possible, more subtle behavior of the ratio $\frac{\mu(x_{1:n})}{\rho(x_{1:n})}$ should be analyzed to obtain conditions both necessary and sufficient for prediction. This should give rise to an abstract characterization of classes of measures for which a measure satisfying such conditions for all measures in the class exists; that is, to a description of classes of measures for which prediction is possible. It is expected that such characterization will naturally lead to a construction of a predictor as well — perhaps in form of a Bayesian integral. The next step will be to extend this approach to the task of active learning.

References

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

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

[3] M. Jackson, Ehud Kalai, and Rann Smorodinsky. Bayesian Representation of Stochastic Processes under Learning: de Finetti Revisited. *Econometrica*, 67(4):875–794, 1999.

[4] B. Ryabko. Prediction of random sequences and universal coding. *Problems of Information Transmission*, 24(2):87–96, 1988.

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

[6] R. J. Solomonoff. A formal theory of inductive inference: Part 1 and 2. *Inform. Control*, 7:1–22, 224–254, 1964.

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

[8] A. K. Zvonkin and L. A. Levin. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Mathematical Surveys*, 25(6):83–124, 1970.