

# Some sufficient conditions on an arbitrary class of stochastic processes for the existence of a predictor.

Daniil Ryabko

► **To cite this version:**

Daniil Ryabko. Some sufficient conditions on an arbitrary class of stochastic processes for the existence of a predictor.. 19th International Conference on Algorithmic Learning Theory, ALT 2008, Oct 2008, Budapest, Hungary. pp.169-182, 10.1007/978-3-540-87987-9\_17 . inria-00347706

**HAL Id: inria-00347706**

**<https://hal.inria.fr/inria-00347706>**

Submitted on 16 Dec 2008

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Some sufficient conditions on an arbitrary class of stochastic processes for the existence of a predictor.

Daniil Ryabko daniil@ryabko.net

INRIA Lille–Nord Europe, France

**Abstract.** We consider the problem of sequence prediction in a probabilistic setting. Let there be given a class  $\mathcal{C}$  of stochastic processes (probability measures on the set of one-way infinite sequences). We are interested in the question of what are the conditions on  $\mathcal{C}$  under which there exists a predictor (also a stochastic process) for which the predicted probabilities converge to the correct ones if any of the processes in  $\mathcal{C}$  is chosen to generate the data. We find some sufficient conditions on  $\mathcal{C}$  under which such a predictor exists. Some of the conditions are asymptotic in nature, while others are based on the local (truncated to first observations) behaviour of the processes. The conditions lead to constructions of the predictors. In some cases we also obtain rates of convergence that are optimal up to an additive logarithmic term.

## 1 Introduction

Given a finite sequence  $x_1, \dots, x_n$  of observations  $x_i \in \mathcal{X}$ , where  $\mathcal{X}$  is a finite set, we want to predict what are the probabilities of observing  $x_{n+1} = x$  for each  $x \in \mathcal{X}$ . It is assumed that the sequence is generated by some unknown stochastic process  $\mu$ , a probability measure on the set of one-way infinite sequences  $\mathcal{X}^\infty$ . The goal is to have a predictor such that the difference between the predicted and correct probabilities goes to zero (in some sense). In general this goal is impossible to achieve if nothing is known about the measure  $\mu$  generating the sequence. In other words, one cannot have a predictor whose error goes to zero for any measure  $\mu$ . However, if  $\mu$  is known to belong to a certain class  $\mathcal{C}$  of measures some well-known results establish the existence of a predictor.

In particular, the Laplace measure

$$\rho_L(x_{n+1} = a | x_1, \dots, x_n) = \frac{\#\{i \leq n : x_i = a\} + 1}{n + |\mathcal{X}|}$$

predicts any Bernoulli i.i.d. process, that is, predicted probabilities converge to “true” probabilities if the measure generating the sequence is a Bernoulli i.i.d. process. Based on similar ideas a predictor can be constructed for the class of all  $k$ -order Markov measures, and, moreover, such predictors can be combined [11] to form a predictor for the class of all stationary processes over  $\mathcal{X}^\infty$ . As another example, one can construct a predictor for any given countable class of measures, as shown by Solomonoff’s construction of a predictor [16] for the class of all semi-computable measures.

Thus there are examples of classes of processes for which a predictor is known to exist. These examples cover some cases interesting theoretically or important from the application point of view. On the other hand, a trivial negative example is a class of all deterministic sequences (that is, each measure in the class produces a certain sequence of outcomes with probability 1) for which a predictor does not exist: for any predictor there is a measure in the class (a deterministic sequence) on which the predicted probabilities differ from the “true” ones by at least  $1/2$  on every step.

**The question we are addressing** in this work is: in general, for which classes  $\mathcal{C}$  of stochastic processes there exists a predictor that predicts every measure in the class?

**Motivation.** The importance of this question stems primarily from the fact that, interesting as the studied cases are, their motivation originally comes either from specific applications or from theoretical attractiveness of the corresponding assumptions. Since new and new applications for the problem of sequence prediction constantly come to existence, known theoretical models can be unsuitable for some of them. For example, stationary processes may model well some physical phenomena but may be less suited for analysis of DNA sequences. If one had a tool to check feasibility of different theoretical assumptions (that is, to check whether there is a predictor that predicts every process satisfying these assumptions) one could use it to find a better model for each specific application.

**Prior work.** Apart from the results on the examples of classes  $\mathcal{C}$  mentioned above (i.i.d., finite-memory, stationary, computable), this general question (at least for sequence prediction) has received little attention. A related question has been addressed in [8]: Whether, given a class of measures  $\mathcal{C}$  and a prior (“meta”-measure)  $\lambda$  over this class of measures, the conditional probabilities of a Bayesian mixture of the class  $\mathcal{C}$  w.r.t.  $\lambda$  converge to the true  $\mu$ -conditional probabilities (weakly merge, in terminology of [8]) for  $\lambda$ -almost any measure  $\mu$  in  $\mathcal{C}$ . The answer found in [8] is a set of necessary and sufficient conditions on the measure given by the mixture of the class  $\mathcal{C}$  w.r.t.  $\lambda$  under which prediction is possible. The major difference from the general question 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.

Another related question is formulated as a question about two individual measures rather than a class of measures and a predictor. Namely, one can ask under which conditions one stochastic process predicts another. In [3] it was shown that if one measure is absolutely continuous with respect to another, than the latter predicts the former (the conditional probabilities converge in a very strong sense). In [13] a weaker form of convergence of probabilities (in particular, convergence of expected average KL divergence) is obtained under weaker assumptions.

**Measuring prediction quality.** As it was mentioned, we are interested in probabilities of observing  $x_{n+1} = x$ ,  $x \in \mathcal{X}$  conditional on  $x_1, \dots, x_n$ . Such conditional probabilities, if specified for every  $x_1, \dots, x_n$  also define a probability measure over  $\mathcal{X}^\infty$ . Thus a predictor (for a class of stochastic processes) is also a stochastic process. The quality of prediction is measured as the discrepancy between the predicted and “true” conditional probabilities. In this work we are mainly considering the Kullback-Leibler

divergence between conditional probabilities, averaged over time, which is either required to converge to zero in expectation over  $x_1, \dots, x_n$  (expectation being taken with respect to the “true” measure generating the sequence), or with probability 1 (again with respect to the measure generating the sequence). Thus, we are interested in the conditions on a class  $\mathcal{C}$  of measures, under which there exists a measure  $\rho_{\mathcal{C}}$  such that average KL divergence between  $\rho$  and  $\mu$  conditional probabilities goes to zero for every  $\mu \in \mathcal{C}$ , in  $\mu$ -expectation or with  $\mu$ -probability 1.

**The results.** In the present work we exhibit some sufficient conditions on the class  $\mathcal{C}$  under which this is possible; none of these conditions relies on parametrization of any kind. The conditions presented are of two types: conditions on asymptotic behaviour of measures in  $\mathcal{C}$ , and on their local (restricted to first  $n$  observations) behaviour. Conditions of the first type concern separability of  $\mathcal{C}$  with respect to the expected average KL divergence. We show that such separability is sufficient for the existence of a predictor.

The conditions of the second kind concern the “capacity” of the set  $\mathcal{C}^n := \{\mu^n : \mu \in \mathcal{C}\}$  where  $\mu^n$  is the measure  $\mu$  restricted to the first  $n$  observations. Intuitively, if  $\mathcal{C}^n$  is small in some sense then prediction is possible. We measure the capacity in two ways. The first way is to find the maximum probability given to each sequence  $x_1, \dots, x_n$  by some measure in the class, and then take a sum over  $x_1, \dots, x_n$ . Denoting the obtained quantity  $c_n$ , one can show that it grows polynomially in  $n$  for some important classes of processes, such as i.i.d. or Markov processes. We show that, in general, if  $c_n$  grows subexponentially then a predictor exists that predicts any measure in  $\mathcal{C}$  in expected average KL divergence. On the other hand, exponentially growing  $c_n$  are not sufficient for prediction. Under slightly stronger conditions on the speed of growth of  $c_n$ , we also establish the existence of a measure that predicts every process  $\mu$  in  $\mathcal{C}$  in average KL divergence with  $\mu$ -probability 1 (rather than in expectation).

A more refined way to measure the capacity of  $\mathcal{C}^n$  is using a concept of channel capacity from information theory, which was developed for a closely related problem of finding optimal codes for a class of sources. We extend corresponding results from information theory to show that sublinear growth of channel capacity is sufficient for the existence of a predictor, in the sense of expected average divergence. Moreover, the obtained bounds on the divergence are optimal up to an additive logarithmic term.

## 2 Preliminaries

We consider stochastic processes (probability measures) 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 symbol  $\mu$  is reserved for the “true” measure generating the sequence. We use  $\mathbf{E}_\nu$  for expectation with respect to a measure  $\nu$  and simply  $\mathbf{E}$  for  $\mathbf{E}_\mu$  (expectation with respect to the “true” measure generating the sequence).

To measure the quality of prediction we will mainly use quantities which are based on the Kullback-Leibler (KL) divergence. For two probability distributions  $\nu_1$  and  $\nu_2$  on a finite set  $\mathcal{X}$  the *KL divergence*  $d(\nu_1, \nu_2)$  is defined as

$$d(\nu_1, \nu_2) = \sum_{x \in \mathcal{X}} \nu_1(x) \log \frac{\nu_1(x)}{\nu_2(x)}. \quad (1)$$

The quality of prediction can be measured as time-average KL divergence between forecast and true probabilities. Thus for a sequence  $(x_1, \dots, x_n) \in \mathcal{X}^n$  the *average KL divergence* between  $\mu$  and  $\rho$  is defined as

$$\bar{d}_n(\mu, \rho, x_1, \dots, x_n) = \frac{1}{n} \sum_{t=1}^n d_t(\mu(\cdot|x_1, \dots, x_{t-1})\rho(\cdot|x_1, \dots, x_{t-1})), \quad (2)$$

where  $\mu(\cdot|x_1, \dots, x_{t-1})$  is the probability distribution of the  $t$ th member of the sequence conditional on  $x_1, \dots, x_{t-1}$ .

We say that  $\rho$  predicts  $\mu$  in *average KL divergence* if

$$\bar{d}_n(\mu, \rho|x_1, \dots, x_n) \rightarrow 0 \text{ } \mu\text{-a.s.},$$

and  $\rho$  predicts  $\mu$  in *expected average KL divergence* if

$$\mathbf{E}_\mu \bar{d}_n(\mu, \rho|x_1, \dots, x_n) \rightarrow 0.$$

We also define *asymptotic expected KL divergence* between measures  $\mu_1$  and  $\mu_2$  as

$$D(\mu_1, \mu_2) = \limsup_{n \rightarrow \infty} \mathbf{E}_{\mu_1} \bar{d}_n(\mu_1, \mu_2|x_1, \dots, x_{n-1}).$$

We will often omit the argument  $x_1, \dots, x_n$  from our notation.

### 3 Main results

**Asymptotic conditions.** Call a class  $\mathcal{C}$  of stochastic processes *separable with respect to (asymptotic expected KL divergence)  $D$*  if there is a countable set  $M \subset \mathcal{C}$  with the following property: For every  $\mu \in \mathcal{C}$  and every  $\varepsilon > 0$  there is  $\mu_\varepsilon \in M$  such that  $D(\mu, \mu_\varepsilon) \leq \varepsilon$ .

**Theorem 1.** *If  $\mathcal{C}$  is separable with respect to  $D$  then there exists a measure  $\rho$  such that  $\rho$  predicts every  $\mu \in \mathcal{C}$  in expected average KL divergence*

$$\mathbf{E}_\mu \bar{d}_n(\mu, \rho, x_1, \dots, x_n) \rightarrow 0$$

for every  $\mu \in \mathcal{C}$ .

*Proof.* Let  $w_k, k \in \mathbb{N}$  be a sequence of positive reals that sum to 1, e.g.  $w_k = 2^{-k}$ . Since the set  $M$  is countable we can introduce  $\mu_i, i \in \mathbb{N}$  such that  $M = \{\mu_i : i \in \mathbb{N}\}$ . Define the predictor  $\rho$  as  $\rho = \sum_{i \in \mathbb{N}} w_i \mu_i$ . We have to show that

$$\lim_{n \rightarrow \infty} \mathbf{E}_\mu d_n(\mu, \rho) = 0$$

for every  $\mu \in \mathcal{C}$ . Fix any  $\mu \in \mathcal{C}$  and  $\varepsilon > 0$ . Find  $\mu_k \in M$  such that  $D(\mu, \mu_k) \leq \varepsilon$ . Introduce the symbol  $\mathbf{E}^t$  for  $\mu$ -expectation over  $x_t$  conditional on  $x_1, \dots, x_{t-1}$ . We

have

$$\begin{aligned}
\mathbf{E}_\mu \bar{d}_n(\mu, \rho) &= \frac{1}{n} \mathbf{E} \sum_{t=1}^n \sum_{x_t \in \mathcal{X}} \mu(x_t | x_1, \dots, x_{t-1}) \log \frac{\mu(x_t | x_1, \dots, x_{t-1})}{\rho(x_t | x_1, \dots, x_{t-1})} \\
&= \frac{1}{n} \sum_{t=1}^n \mathbf{E} \mathbf{E}^t \log \frac{\mu(x_t | x_1, \dots, x_{t-1})}{\rho(x_t | x_1, \dots, x_{t-1})} = \frac{1}{n} \mathbf{E} \log \prod_{t=1}^n \frac{\mu(x_t | x_1, \dots, x_{t-1})}{\rho(x_t | x_1, \dots, x_{t-1})} \\
&= \frac{1}{n} \mathbf{E} \log \frac{\mu(x_1, \dots, x_n)}{\rho(x_1, \dots, x_n)} \leq \frac{1}{n} \mathbf{E} \log \frac{\mu(x_1, \dots, x_n)}{w_k \mu_k(x_1, \dots, x_n)} \\
&= \frac{\log w_k^{-1}}{n} + \frac{1}{n} \mathbf{E} \log \frac{\mu(x_1, \dots, x_n)}{\mu_k(x_1, \dots, x_n)} \\
&= \frac{\log w_k^{-1}}{n} + \mathbf{E}_\mu \bar{d}_n(\mu, \mu_k),
\end{aligned}$$

from which we conclude that

$$\limsup_{n \rightarrow \infty} \mathbf{E}_\mu \bar{d}_n(\mu, \rho) \leq \limsup_{n \rightarrow \infty} \mathbf{E}_\mu \bar{d}_n(\mu, \mu_k) \leq \varepsilon.$$

Since this holds for every  $\varepsilon$ , and since KL divergence is always non-negative, we get the statement  $\lim_{n \rightarrow \infty} \mathbf{E}_\mu \bar{d}_n(\mu, \rho) = 0$ .  $\square$

**Example: countable classes.** A trivial but interesting example in which the conditions of Theorem 1 are satisfied is when the class  $\mathcal{C}$  itself is countable. A well-studied case is when  $\mathcal{C}$  is the class of all (semi-)computable measures ([16], see also [7]).

**Example: i.i.d.** Another simple example is given by the class  $\mathcal{C}_B$  of all Bernoulli i.i.d. processes, with  $\mathcal{X} = 0, 1$ , indexed by parameter  $p \in [0, 1]$ ; that is,  $\mu_p(x_n = 0) = p$  for all  $n$  independently of each other. In this case, it is easy to check that the subset of all processes with rational parameters is dense in  $\mathcal{C}_B$  with respect to expected average KL divergence, since  $\bar{d}_n(\mu_p, \mu_q) = d(\mu_p, \mu_q)$  and the latter is continuous in  $p$  and  $q$ .

**Example: Finite-memory, stationary.** The same holds for the class of stationary finite memory processes: each process with memory  $k$  is parametrized by  $|\mathcal{X}|^{k+1}$  parameters — the conditional probabilities of observing  $x_{k+1} = x \in \mathcal{X}$  given  $x_1, \dots, x_k$ . The set of processes with rational values of the parameters is dense with respect to the expected average divergence. Since any stationary ergodic process can be arbitrary well approximated (in the sense of asymptotic expected average KL divergence  $D(\mu, \rho)$ , where  $\limsup$  actually becomes  $\lim$ ) by finite-memory processes, in particular by those with rational parameters, we can conclude that the class of stationary ergodic sources is separable with respect to expected average KL divergence. Thus, applying Theorem 1 we can obtain a different (though based on similar ideas) proof of the result of [11] which says that there exists a predictor for the class of all stationary ergodic processes.

**Conditions based on local behaviour of measures.** Next we provide some sufficient conditions for the existence of a predictor based on local characteristics of the class of measures.

For a class  $\mathcal{C}$  of stochastic processes and a sequence  $(x_1, \dots, x_n) \in \mathcal{X}^n$  introduce the coefficients

$$c_{x_1, \dots, x_n}(\mathcal{C}) = \sup_{\mu \in \mathcal{C}} \mu(x_1, \dots, x_n). \quad (3)$$

Define also the normalizer

$$c_n(\mathcal{C}) = \sum_{(x_1, \dots, x_n) \in \mathcal{X}^n} c_{x_1, \dots, x_n}(\mathcal{C}). \quad (4)$$

A normalized maximum likelihood estimator  $\lambda$  is defined as

$$\lambda_{\mathcal{C}}(x_1, \dots, x_n) = \frac{1}{c_n(\mathcal{C})} c_{x_1, \dots, x_n}(\mathcal{C}). \quad (5)$$

For finite spaces (that is, for fixed  $n$ ) normalized maximum likelihood estimators have been studied in e.g. [15, 2], in the context of information theory. However,  $\lambda_{\mathcal{C}}$  in general do not define a stochastic process over  $\mathcal{X}^\infty$  (they are not consistent for different  $n$ ); thus, in particular, using average KL divergence for measuring prediction quality would not make sense, since  $d_n(\mu(\cdot|x_1, \dots, x_{n-1}), \lambda(\cdot|x_1, \dots, x_{n-1}))$  can be negative.

Yet, by taking an appropriate mixture, it is still possible to construct a predictor (a stochastic process) based on  $\lambda$  that predicts the measures in the class not only in expectation but also with probability 1.

**Theorem 2.** *Suppose that a class  $\mathcal{C}$  of stochastic processes is such that*

$$\log c_n(\mathcal{C}) = o(n). \quad (6)$$

*Then there exists a stochastic process  $\rho$  such that*

$$\mathbf{E}_\mu \bar{d}_n(\mu, \rho, x_1, \dots, x_n) \leq \frac{\log c_n(\mathcal{C})}{n} + O\left(\frac{\log n}{n}\right); \quad (7)$$

*in particular  $\rho$  predicts every  $\mu \in \mathcal{C}$  in expected average KL divergence. If the coefficients  $c_n(\mathcal{C})$  are such that*

$$\sum_{n=1}^{\infty} \frac{\log^2 c_n(\mathcal{C})}{n^2} < \infty \quad (8)$$

*then there exists a stochastic process  $\rho$  that predicts every  $\mu \in \mathcal{C}$  in average KL divergence (with  $\mu$ -probability 1).*

*Proof.* Let  $w := \sum_{k=1}^{\infty} \frac{1}{k^2}$  and let  $w_k := \frac{1}{wk^2}$ . Moreover, define a measure  $\mu_k$  as follows. On first  $k$  steps it is defined as  $\lambda_k$ , and for  $n > k$  it outputs only zeros with probability 1; so,  $\mu_k(x_1, \dots, x_k) = \lambda_{\mathcal{C}}(x_1, \dots, x_k)$  and  $\mu_k(x_n = 0) = 1$  for  $n > k$ .

Finally, let  $\rho = \sum_{k=1}^{\infty} w_k \mu_k$ . We will show that under the conditions of the theorem  $\rho$  has the asserted predictive properties.

For the first statement, we have (similarly to the proof of Theorem 1)

$$\begin{aligned} \mathbf{E}_\mu \bar{d}_n(\mu, \rho) &= \frac{1}{n} \mathbf{E} \log \frac{\mu(x_1, \dots, x_n)}{\rho(x_1, \dots, x_n)} \leq \frac{1}{n} \mathbf{E} \log \frac{\mu(x_1, \dots, x_n)}{w_n \mu_n(x_1, \dots, x_n)} \\ &\leq \frac{1}{n} \log \frac{c_n(\mathcal{C})}{w_n} = \frac{1}{n} (\log c_n(\mathcal{C}) + 2 \log n + \log w). \quad (9) \end{aligned}$$

In order to prove the second statement, we first introduce a short-hand notation  $x_{1..n}$  for  $x_1, \dots, x_n$ . Consider random variables

$$l_n = \log \frac{\mu(x_n | x_{1..n-1})}{\rho(x_n | x_{1..n-1})}$$

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  $x_1, \dots, x_n, \dots$ . Let also  $\bar{m}_n = \frac{1}{n} \sum_{t=1}^n m_t$ . We will show that  $\bar{m}_n \rightarrow 0$   $\mu$ -a.s. and  $\bar{l}_n \rightarrow 0$   $\mu$ -a.s. which implies  $\bar{d}_n \rightarrow 0$   $\mu$ -a.s.

Note that

$$\bar{l}_n = \frac{1}{n} \log \frac{\mu(x_{1..n})}{\rho(x_{1..n})} \leq \frac{\log w_n^{-1} c_n(\mathcal{C})}{n} \rightarrow 0.$$

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 ( $\mu$ -a.s.) bounded from below by a constant, since  $\frac{\rho(x_{1..n})}{\mu(x_{1..n})}$  is a positive  $\mu$ -martingale whose expectation is 1, and so it converges to a finite limit  $\mu$ -a.s. by Doob's submartingale convergence theorem, see e.g. [14, p.508].

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

$$\begin{aligned} m_n &= \log \frac{\mu(x_{1..n})}{\rho(x_{1..n})} - \log \frac{\mu(x_{1..n-1})}{\rho(x_{1..n-1})} - \mathbf{E}^n \log \frac{\mu(x_{1..n})}{\rho(x_{1..n})} + \mathbf{E}^n \log \frac{\mu(x_{1..n-1})}{\rho(x_{1..n-1})} \\ &= \log \frac{\mu(x_{1..n})}{\rho(x_{1..n})} - \mathbf{E}^n \log \frac{\mu(x_{1..n})}{\rho(x_{1..n})}. \end{aligned}$$

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

$$\sum_{n=1}^{\infty} \frac{(\log w_n^{-1} c_n(\mathcal{C}) + f(n))^2}{n^2} < \infty \quad (10)$$

(e.g. choose  $f(n) = \log n$ ). For a sequence of random variables  $\lambda_n$  define

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

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  $\bar{m}_n^+ \rightarrow 0$  we can use the martingale



strong law of large numbers [14, p.501], which states that, for a martingale difference sequence  $\gamma_n$ , if  $\mathbf{E}(n\bar{\gamma}_n)^2 < \infty$  and  $\sum_{n=1}^{\infty} \mathbf{E}\gamma_n^2/n^2 < \infty$  then  $\bar{\gamma}_n \rightarrow 0$  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

$$|m_n^+| \leq \log w_n^{-1} c_n(\mathcal{C}) + f(n)$$

and (10).

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  $\mu$ -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

$\mathbf{E}^n \left( \log \frac{\mu(x_{1..n})}{\rho(x_{1..n})} \right)^{-f} \rightarrow 0$  we write

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

and note that the first term in brackets is bounded from below, and so for the sum in brackets to be less than  $-f(n+1)$  (which is unbounded) the second term  $\log \mu(x_n|x_{1..n})$  has to go to  $-\infty$ , but then the expectation goes to zero since  $\lim_{u \rightarrow 0} u \log u = 0$ .

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

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

$$\sup_{\mu \in \mathcal{C}_B} \mu(x_1, \dots, x_n) = p^k (1-p)^{n-k}$$

where  $k = \#\{i \leq n : x_i = 0\}$  is the number of 0s in  $x_1, \dots, x_n$  and  $p = k/n$ . For the constants  $c_n(\mathcal{C})$  we can get the bound  $c_n(\mathcal{C}) \leq \frac{1}{n+1}$ . In general, for the class  $\mathcal{C}_k$  of processes with memory  $k$  over a finite space  $\mathcal{X}$  we get polynomial  $c_n(\mathcal{C})$  (see e.g. [13]).

Thus, with respect to the finite-memory processes, the conditions of Theorem 2 leave ample space for growth of  $c_n(\mathcal{C})$ : the condition (6) allows any subexponential growth of  $c_n(\mathcal{C})$  and the condition (8) allows for example  $c_n(\mathcal{C}) = 2^{-\sqrt{n}/\log n}$ .

**Example: exponential coefficients are not sufficient.** Observe that the condition (6) cannot be relaxed further, in the sense that exponential coefficients  $c_n$  are not sufficient for prediction. Indeed, for the class of all deterministic processes (that is, each process from the class produces some fixed sequence of observations with probability 1) we have  $c_n = 2^n$ , while obviously for this class a predictor does not exist.

**Optimal rates of convergence.** A natural question that arises with respect to the bound (7) is whether it is optimal (that is, whether it under the conditions formulated. This question is closely related to the optimality of the normalized maximum likelihood estimates used in the construction of the predictor. In general, since such estimates are not optimal neither are the rates of convergence in (7). To obtain (close to) optimal rates one has to consider a different measure of capacity.

To do so, we make the following connection to a problem in information theory. For a class  $\mathcal{C}$  of measures we are interested in a predictor that has small (or minimal) worst-case (with respect to the class  $\mathcal{C}$ ) probability of error. Thus, we are interested in the quantity

$$\inf_{\rho} \sup_{\mu \in \mathcal{C}} D(\mu, \rho), \quad (11)$$

where the infimum is taken over all stochastic processes  $\rho$  and  $D$  is the asymptotic expected average KL divergence. (In particular, we are interested in the conditions under which the quantity in (11) equals zero.) This problem has been studied for the case when the probability measures are over a finite set  $\mathcal{X}$ , and  $D$  is replaced simply by the KL divergence  $d$  between the measures. Thus, the problem is to find the probability measure  $\rho$  (if it exists) on which the following minimax is attained

$$R(A) := \inf_{\rho} \sup_{\mu \in A} d(\mu, \rho). \quad (12)$$

This problem is closely related to the problem of finding the best code for the class of sources  $A$ , which was its original motivation. The normalized maximum likelihood distribution considered above does not in general lead to the optimum solution for this problem. The optimum solution is obtained through the result that relates the minimax (12) to the so-called channel capacity. For a set  $A$  of measures on a finite set  $\mathcal{X}$  the *channel capacity* of  $A$  is defined as

$$C(A) := \sup_P \sum_{\mu} P(\mu) d(\mu, \rho_P) \quad (13)$$

where  $P$  ranges over all probability distributions over all finite subsets of  $A$  and  $\rho_P = \sum_{\mu} P(\mu)\mu$ . It is shown in [10, 5] that  $C(A) = R(A)$ , thus reducing the problem of finding a minimax to an optimization problem. Moreover, Arimoto-Blahut algorithm [1, 4] is used to approximate  $C(A)$  numerically and solve the optimization problem for the important case when  $A$  is the convex hull of a finite set. For probability measures over infinite spaces this result ( $R(A) = C(A)$ ) was generalized in [6], but the divergence between probability distributions is measured by KL divergence (and not average

KL divergence), which gives infinite  $R(A)$  for most of the cases interesting from the sequence prediction point of view (e.g. for the class of Bernoulli i.i.d. processes).

However, truncating measures in a class  $\mathcal{C}$  to the first  $n$  observations, we can use the results about channel capacity to analyze the predictive properties of the class. Moreover, the rates of convergence that can be obtained along these lines are close to optimal.

**Theorem 3.** *Let  $\mathcal{C}$  be a class of measures over  $\mathcal{X}^\infty$  and  $\mathcal{C}^n$  be the class of measures from  $\mathcal{C}$  restricted on  $\mathcal{X}^n$ . There exists a measure  $\rho$  such that*

$$\mathbf{E}_\mu \bar{d}_n(\mu, \rho, x_1, \dots, x_n) \leq \frac{\log C_n(\mathcal{C})}{n} + O\left(\frac{\log n}{n}\right). \quad (14)$$

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

$$\mathbf{E}_\mu \bar{d}_n(\mu, \rho, x_1, \dots, x_n) \geq \frac{\log C_n(\mathcal{C})}{n} - \varepsilon.$$

*Proof.* As shown in [5], for each  $n$  there exists a sequence  $\nu_k^n$ ,  $k \in \mathbb{N}$  of measures on  $\mathcal{X}^n$  such that

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

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

$$\left| \sup_{\mu^n \in \mathcal{C}^n} d(\mu^n, \nu_{k_n}^n) - C(\mathcal{C}^n) \right| \leq 1/n.$$

Define the measure  $\rho_n$  as follows. On first  $n$  symbols it coincides with  $\nu_{k_n}^n$  and  $\rho(x_m = 0) = 1$  for  $m > n$ . Finally, set  $\rho = \sum_{n=1}^{\infty} w_n \rho_n$ , where  $w_k = \frac{1}{w n^2}$ ,  $w = \sum_{n=1}^{\infty} \frac{1}{n^2}$ . We have to show that  $\mathbf{E}_\mu \bar{d}_n(\mu, \rho) = 0$  for every  $\mu \in \mathcal{C}$ . Indeed,

$$\begin{aligned} \mathbf{E}_\mu \bar{d}_n(\mu, \rho) &= \frac{1}{n} \mathbf{E}_\mu \log \frac{\mu(x_{1..n})}{\rho(x_{1..n})} \\ &\leq \frac{\log w_k^{-1}}{n} + \frac{1}{n} \mathbf{E}_\mu \log \frac{\mu(x_{1..n})}{\rho_n(x_{1..n})} \leq \frac{\log w + 2 \log n}{n} + \frac{1}{n} d(\mu^n, \rho_n) \\ &\leq o(1) + \frac{C(\mathcal{C}^n)}{n} + \frac{1}{n} = o(1). \end{aligned} \quad (15)$$

The second statement follows from the fact [10, 5] that  $C(\mathcal{C}^n) = R(\mathcal{C}^n)$  (cf. (12)).  $\square$

Thus if the channel capacity  $C(\mathcal{C}^n)$  grows sublinearly a predictor can be constructed for the class of processes  $\mathcal{C}$ . In this case the problem of constructing the predictor is reduced to finding the channel capacities for different  $n$  and finding the corresponding measures on which it is attained or approached.

As an **example** we can mention, again, the class of all Bernoulli processes, whose channel capacity  $C(\mathcal{C}_B^n)$  is  $O(\log n)$ , see e.g. [9].

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

## 4 Discussion

As far as **algorithmic realizability** of the predictors proposed is concerned, we should first note that when an input parameter of an “algorithm” is an arbitrary class of stochastic processes, one can hardly talk about algorithms for real computers. Rather, the predictors constructed have to be regarded as *reductions* of the problem of finding a predictor for a given class of stochastic processes to the conceptually much easier problems of approximating certain suprema and infinite sums. Here an analogy can be made with the classification problem. In certain cases the problem of finding a good classifier can be reduced to the problem of finding a classifier from a given class that best fits the data (minimizes empirical risk [17]). Conceptually this is a much simpler problem; however, in some cases it can be intractable (see e.g. [12]). In general, for each particular class of classifiers a separate algorithm should be constructed to find efficiently a classifier that fits the data. Indeed, efficient solutions (such as support vector machines [17]) exist for many important cases.

Returning to our problem, Theorem 1 states that if in a class  $\mathcal{C}$  of measures there is a countable dense subset  $M$ , then a predictor can be constructed whose average expected error goes to zero. Moreover, such a predictor can be obtained as a weighted sum of measures from  $M$  (with any positive weights that sum to 1). Thus the problem of finding a predictor is reduced to two (simpler) problems: finding a dense subset and taking an infinite sum. We can further show that in some cases the latter problem can be reduced to a version of the former, that is, it is not necessary to take an infinite sum if one can find a finite  $\varepsilon$ -net.

**Proposition 1.** *Let a class  $\mathcal{C}$  of stochastic processes be such that for some  $\varepsilon$  there exists a subset  $M_\varepsilon \subset \mathcal{C}$  with the following property. For any  $\mu \in \mathcal{C}$  there exists  $\mu' \in M_\varepsilon$  such that  $D(\mu, \mu') \leq \varepsilon$ . Then for the measure*

$$\rho = \sum_{\nu \in M_\varepsilon} w_\nu \nu,$$

where  $w_\nu$  are any positive reals (that sum to 1), we have  $D(\mu, \rho) \leq \varepsilon$  for any  $\mu \in \mathcal{C}$ .

This statement can be proven in exactly the same way as Theorem 1.

The predictors constructed in the proofs of Theorems 2 and 3 also involve summation over an infinite set. However, in these cases it is immediately apparent from the constructions of the predictors that for prediction on  $n$ th step it is sufficient to take sums up to  $n$ , and the bounds on expected average divergence (7) and (14) still hold. Thus the problem of finding a predictor is reduced to the problem of approximating a finite number of suprema.

Namely, for the case of normalized maximum likelihood predictor of Theorem 2 the quantity (3) have to be evaluated for each  $n$  and each  $x_1, \dots, x_n$ . For the predictor based on channel capacity one has to find the measure on which channel capacity (13) is attained (possibly up to a certain  $\varepsilon_n$ ) for each  $n$ . As it was mentioned, for the latter problem one can use Arimoto-Blahut algorithm [1, 4] in the case when the set  $\mathcal{C}^n$  is a convex hull of a finite number of measures. We also have to note that the requirement of convexity is not really a restriction since a predictor for a set of measures is also a

predictor for its convex hull, and so we can always consider convex hulls of classes of predictors rather than classes themselves.

One more question we discuss is **other possible ways of measuring the quality of prediction**. In this paper we were considering KL divergence (1) averaged over time (2), and have developed predictors on which this divergence tends to zero either in expectation or with probability 1. Other possible ways of measuring divergence include absolute distance

$$a_n(\mu, \rho|x_{1..n-1}) = \sum_{x \in \mathcal{X}} |\mu(x_n = x|x_{1..n-1}) - \rho(x_n = x|x_{1..n-1})|,$$

squared absolute distance

$$s_n(\mu, \rho|x_{1..n-1}) = \sum_{x \in \mathcal{X}} (\mu(x_n = x|x_{1..n-1}) - \rho(x_n = x|x_{1..n-1}))^2,$$

and Hellinger distance

$$h_n(\mu, \rho|x_{1..n-1}) = \sum_{x \in \mathcal{X}} (\sqrt{\mu(x_n = x|x_{1..n-1})} - \sqrt{\rho(x_n = x|x_{1..n-1})})^2.$$

Analogously with the average KL divergence (2) we can define average absolute distance  $\bar{a}_n$ , average squared absolute distance  $\bar{s}_n$  and average Hellinger distance  $\bar{h}_n$ . Using Pinsker's inequality  $a_t^2 \leq 2d_t$  one can easily show that all convergence results and upper bounds stated in terms of KL divergence also hold for the measures of divergence just introduced (see e.g. Lemma 1 of [13] for details).

**Proposition 2.** *The statements concerning convergence and upper of Theorems 1, 2 and 3 hold true if average KL distance  $\bar{d}_n$  is replaced by either of the following: average absolute distance  $\bar{a}_n$ , average squared absolute distance  $\bar{s}_n$  and average Hellinger distance  $\bar{h}_n$ .*

Another interesting problem would be to investigate a stronger notion of predictive quality: without averaging over time. For example, under which conditions on a class  $\mathcal{C}$  of measures there exist a predictor  $\rho$  for which

$$d_n(\mu, \rho|x_1, \dots, x_{n-1})$$

goes to zero with  $\mu$  probability 1 for every  $\mu$ .

These questions, along with the problem of finding efficient algorithmic solutions for cases of practical interest, constitute an agenda for further investigation.

## References

1. S. Arimoto, An algorithm for computing the capacity of arbitrary discrete memoryless channels, *IEEE Transactions on Information Theory* IT-18: 14–20, 1972.
2. Qun Xie, A. Barron, Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Transactions on Information Theory* 46(2): 431–445.

3. D. Blackwell and L. Dubins. Merging of opinions with increasing information. *Annals of Mathematical Statistics*, 33:882–887, 1962.
4. R. Blahut, Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18 (4): 460–473, 1972.
5. R. Gallager, Source Coding With Side Information and Universal Coding, M.I.T. LIDS-P-937, 1976 (revised 1979).
6. D. Haussler, A General Minimax Result for Relative Entropy. *IEEE Transactions on Information Theory* 43 (4): 1276–1280, 1997.
7. M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005.
8. M. Jackson, E. Kalai, and R. Smorodinsky. Bayesian representation of stochastic processes under learning: de Finetti revisited. *Econometrica*, 67(4):875–794, 1999.
9. R. Krichevsky, Universal Compression and Retrieval, Kluwer Academic Publishers, 1993.
10. B. Ryabko. Coding of a source with unknown but ordered probabilities. *Problems of Information Transmission* 15 (2):134–138, 1979.
11. B. Ryabko. Prediction of random sequences and universal coding. *Problems of Information Transmission*, 24:87–96, 1988.
12. D. Ryabko, On sample complexity for computational classification problems, *Algorithmica*, 49 (1): 69–77, 2007.
13. D. Ryabko, M. Hutter. Predicting Non-Stationary Processes, *Applied Mathematics Letters* 21(5): 477–482, 2008.
14. A. Shiryaev. *Probability*. Springer, 1996.
15. Yu. Shtarkov, Universal sequential coding of single messages, *Problems of Information Transmission* , 23: 3-17, 1988.
16. R. J. Solomonoff. Complexity-based induction systems: comparisons and convergence theorems. *IEEE Trans. Information Theory*, IT-24:422–432, 1978.
17. V. Vapnik, *Statistical Learning Theory*, New York etc.: John Wiley , Sons, Inc. 1998.