Proceedings of Machine Learning Research 76:1–15, 2017 Algorithmic Learning Theory 2017

Universality of Bayesian mixture predictors

Daniil Ryabko DANIIL@RYABKO.NET
INRIA, 40 avenue de Halley, 59650 Villeneuve d'Ascq, France

Editors: Steve Hanneke and Lev Reyzin

Abstract

The problem is that of sequential probability forecasting for discrete-valued time series. The data is generated by an unknown probability distribution over the space of all one-way infinite sequences. It is known that this measure belongs to a given set $\mathcal{C}$, but the latter is completely arbitrary (uncountably infinite, without any structure given). The performance is measured by asymptotic average log loss. In this work it is shown that the minimax asymptotic performance is always attainable, and it is attained by a Bayesian mixture over countably many measures from the set $\mathcal{C}$. This was previously only known for the case when the best achievable asymptotic error is 0. The new result can be interpreted as a complete-class theorem for prediction. It also contrasts previous results that show that in the non-realizable case all Bayesian mixtures may be suboptimal. This leads to a very general conclusion concerning model selection for a problem of sequential inference: it is better to take a model large enough to make sure it includes the process that generates the data, even if it entails positive asymptotic average loss, for otherwise any combination of predictors in the model class may be useless.

1. Introduction

Given a sequence $x_1, \dots, x_n$ of observations $x_i \in \mathcal{X}$, where $\mathcal{X}$ is a finite set, we want to predict the probabilities of observing $x_{n+1} = x$ for each $x \in \mathcal{X}$, before x_{n+1} is revealed, after which the process continues sequentially. The sequence $x_1, \dots, x_n, \dots$ is generated by some unknown stochastic process μ, a probability measure on the space of one-way infinite sequences $\mathcal{X}^\infty$. Further, a set $\mathcal{C}$ of such measures is given, and it is known that $\mu \in \mathcal{C}$. The set $\mathcal{C}$ can be thought of as the set of models, experts, or the set of strategies of the adversary (a.k.a. Nature). The requirement that the true measure μ that generates the data is in $\mathcal{C}$ means that we are in the "realizable" case of the problem (in other words, there is at least one expert that is optimal).

Unlike most of the literature on the subject, which assumes that the set $\mathcal{C}$ is parametrized and endowed with some structure, here we would like to treat the problem in full generality and thus shall not make any assumptions whatsoever on the set $\mathcal{C}$ or its elements. Note that making even such innocuous-looking assumptions on the set $\mathcal{C}$ as are the common topological ones, such as local compactness, separability, tightness, not to mention stronger assumptions involving the existence of densities or smoothness, implicitly gives the problem a structure (e.g., a topology in which the assumption is fulfilled) that in itself constitutes a large part of the solution. Here we are interested rather in the fundamental question of which principles to use when choosing a model for a problem, and thus would like not to

make any assumptions at all (not even measurability). It is also worth reiterating that the measures in $\mathcal{C}$ are not required to be i.i.d., finite-memory, mixing, etc.

We are interested in the question whether it is possible to attain the minimax optimal asymptotic performance by using a combination of measures in $\mathcal{C}$ as a predictor. A combination is a measure of the form $\nu = \int_{\mathcal{C}} dW$ where W is some measure over $\mathcal{C}$ (or, more generally, over a measurable subset of $\mathcal{C}$; recall that $\mathcal{C}$ itself is not required to be measurable). The measure W can be thought of as a prior distribution, and prediction is then by evaluating the posterior $\nu(\cdot|x_1, \dots, x_n)$ on the observed sequence $x_1, \dots, x_n$. In other words, we are asking whether it is possible to achieve optimal prediction with a Bayesian predictor with some prior (that is, whether there exists such a prior); however, we are not interested in optimality with prior probability 1, but rather in the minimax (worst-case) asymptotic optimality of such predictors. The answer we obtain is positive.

Thus, **the main result** is that the minimax asymptotic performance is always attainable and it is attained by a combination of countably many measures from the set $\mathcal{C}$. Note that this statement is completely assumption-free: not even measurability of $\mathcal{C}$ is required.

Previously, this result was only established (Ryabko, 2010) under the assumption that there is a predictor whose error is asymptotically 0 on any measure $\mu \in \mathcal{C}$, that is, the minimax asymptotic error is 0. Here we get rid of this (last) assumption. Examples of cases where the minimax asymptotic error is greater than 0 are easiest to come by if we suppose that some aspects of the process are completely arbitrary. The easiest example is when nothing is predictable: the data is an arbitrary deterministic sequence. This example gives the maximal possible worst-case asymptotic error of $\log |\mathcal{X}|$. A more meaningful example is that of processes with (frequent) abrupt changes: between the changes, the distribution belongs to some (nice) given family (e.g., i.i.d. Bernoulli trials) but when the changes occur is not known, and a change is to an arbitrary distribution in the family. This example is considered in more detail in Section 5.

Moreover, the case when the best possible asymptotic error is greater than 0 is particularly important in light of recent results achieved in the non-realizable case, that is, when the measure μ generating the data does not have to belong to $\mathcal{C}$. In this case, one is interested in the regret with respect to $\mathcal{C}$, that is, the performance of the predictor minus the best performance of all the measures in $\mathcal{C}$ on the given μ. Ryabko (2016) shows that, in the non-realizable case, it can happen that the best regret a predictor can achieve with respect to a set $\mathcal{C}$ is zero, but any Bayesian mixture predictor has regret bounded away from 0 by a large constant (see Section 6 for a precise formulation). In other words, the experts in $\mathcal{C}$ are useless: one can do as well as any of them, but not by combining them. Note that any such set $\mathcal{C}$ has to be uncountable, which brings it out of the traditional expert advice settings (a survey of which can be found in Cesa-Bianchi and Lugosi, 2006).

Putting these results together, we reach the following fundamental **recommendation for choosing a model** for sequential data:

Better take a model large enough to make sure it includes the process that generates the data, even if it makes the worst-case asymptotic error larger than zero, for otherwise any combination of predictors in the model class may be useless.

For the initiated reader this result has a distinct decision-theoretic flavour to it. Indeed, as we explain in Section 4, it provides (a strong form) of the the complete-class theorem for the problem of sequential prediction, as well as a partial form of the minimax theorem.

Related work. The literature on (nonparametric) sequential prediction is huge, and we do not attempt to provide an adequate survey here. Some pointers to climb references to and from in different branches of science are: (Cesa-Bianchi and Lugosi, 2006) for the expert advice setting (machine learning side), (Kalai and Lehrer, 1994; Noguchi, 2015) for the non-parametric Bayesian approach (econometrics side; most results with prior probability 1), (Ryabko, 1988; Morvai et al., 1997; Gyorfi and Ottucsak, 2007) for predicting stationary ergodic time series (perhaps the largest class considered in statistics; non-parametric statistics/information theory side), (Solomonoff, 1978; Hutter, 2005) for predicting computable measures. The study of the realizable and non-realizable sides of the prediction problem together in the setting considered here has been initiated by Ryabko (2011) that also poses the question that Ryabko (2016) resolves.

To the author's knowledge, this is the first work to consider the general case case when the best achievable asymptotic regret is greater than 0. One specific example that was considered before is that of processes with abrupt changes mentioned above. Willems (1996) considers the case when the processes between the changes are i.i.d., but the method proposed is general. It is, in fact, a Bayesian construction, where the prior is over all possible sequences of changes. The goal is to minimize the regret with respect to the predictor that knows the sequence of changes (but not the distributions); the best achievable asymptotic regret was not considered directly. Subsequent work on this problem is largely devoted to computational considerations; see (Gyorgy et al., 2012) and references therein.

Related decision-theoretic results concern the setting of the problem for "predicting" just one (the first) symbol of the sequence. For KL divergence (expected log loss) these results include (Ryabko, 1979; Gallager, 1976 (revised 1979; Haussler, 1997); a variety of generalizations to other losses is presented in (Grünwald and Dawid, 2004).

2. Preliminaries

Let $\mathcal{X}$ be a finite set. The notation $x_{1..n}$ is used for $x_1, \dots, x_n$. We consider (probability) measures on $(\mathcal{X}^\infty, \mathcal{F})$, where $\mathcal{F}$ is the usual Borel sigma-field. For a finite set A denote $|A|$ its cardinality. We use $\mathbf{E}_\mu$ for expectation with respect to a measure μ.

For two measures μ and ρ introduce the *expected cumulative Kullback-Leibler divergence (KL divergence)* as

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

In words, we take the expected (over data) average (over time) KL divergence between μ- and ρ-conditional (on the past data) probability distributions of the next outcome; and this gives simply the μ-expected log-ratio of the likelihoods.

Definition 1 *Define the asymptotic average KL loss of ρ on μ as*

$$D(\mu,\rho) := \limsup_{n\to\infty} \frac{1}{n} d_n(\mu,\rho). \tag{2}$$

For a set $\mathcal{C}$ of measures define

$$D(\mathcal{C},\rho) := \sup_{\mu\in\mathcal{C}} D(\mu,\rho).$$

The main quantity of interest is the following minimax loss:

Definition 2 *For a set $\mathcal{C}$ of measures define*

$$V_{\mathcal{C}} := \inf_{\rho\in\mathcal{P}} D(\mathcal{C},\rho) = \inf_{\rho\in\mathcal{P}} \sup_{\mu\in\mathcal{C}} D(\mu,\rho), \tag{3}$$

where the infimum is taken over the set $\mathcal{P}$ of all probability measures on $(\mathcal{X}^\infty,\mathcal{F})$.

Thus, $V_{\mathcal{C}}$ is the minimax loss for the set $\mathcal{C}$ of strategies of the Nature and the unrestricted set of Statistician's strategies.

3. Main result

The main result shows that the minimax loss is always achievable and is achieved by a convex combination (a Bayesian mixture) of measures in $\mathcal{C}$ — without any assumptions on $\mathcal{C}$. Moreover, for any predictor ρ there is a convex combination of measures in $\mathcal{C}$ that is as good as ρ.

Theorem 3 *For any set $\mathcal{C}$ of probability measures on $(\mathcal{X}^\infty,\mathcal{F})$, there exist a sequence of measures $\mu_k\in\mathcal{C}$, $k\in\mathbb{N}$ and a sequence of real weights $v_k>0, k\in\mathbb{N}$ whose sum is 1, such that for the measure $\varphi:=\sum_{k\in\mathbb{N}} v_k\mu_k$ we have*

$$D(\mathcal{C},\varphi) = V_{\mathcal{C}}.$$

Moreover, for every measure ρ there exists a predictor φ of the form above such that $D(\mu,\varphi)\le D(\mu,\rho)$ for all $\mu\in\mathcal{C}$.

Before giving the proof, we present informally some ideas behind it. Imagine first that for the set $\mathcal{C}$ we already knew a predictor ρ that attains the value $V_{\mathcal{C}}$. Imagine furthermore, that for each $\mu\in\mathcal{C}$ the limit $\lim_{n\to\infty} -\frac{1}{n}\log\frac{\rho(x_{1..n})}{\mu(x_{1..n})}$ exists for μ-almost all $\mathbf{x}=x_1,\dots,x_n,\dots\in\mathcal{X}^\infty$. Then we could define ($\mu$-almost everywhere) the function $f_\mu(\mathbf{x})$ whose value at $\mathbf{x}$ equals this limit. Let us call it the "log-density" function. (The reader can recognize behind the log the expression $\lim_{n\to\infty}\frac{\rho(x_{1..n})}{\mu(x_{1..n})}$ that defines the familiar densities.) Furthermore, nothing forbids us to imagine that this log-density is measurable. What we would be looking for thence is to find a countable dense subset of the set of log-densities of all measures from $\mathcal{C}$. The measures μ corresponding to each log-density in this countable set would then constitute the sequence whose existence the theorem asserts. To find such a dense countable subset we could employ a standard procedure: approximate all log-densities by step functions with finitely many steps. The main technical argument is then to show that, for each level of the step functions, there are not too many of these functions whose steps are concentrated on different sets of non-negligible probability, for otherwise the requirement that ρ attains $V_{\mathcal{C}}$ would be violated. Here "not too many" means exponentially many with the right exponent (the one corresponding to the step of the step-function with which we approximate the density), and "non-negligible probability" means a probability bounded away (in n) from 0. Getting back to reality, we cannot say anything about the existence of the limits. What we do instead is use the step-functions approximation at each time step n. Since there are only countably many time steps, the result is still a countable set of measures μ from $\mathcal{C}$.

Finally, we are not given a measure ρ that attains the value $V_{\mathcal{C}}$; so we find a sequence of predictors ρ_{ε_n} that approach this value instead, and perform the procedure above for each ρ_{ε_n}.

It is worth noting that the proof that Ryabko (2010) obtains for the special case $V_{\mathcal{C}} = 0$ does not directly generalize. In fact, tidying up the constants in that proof, one can only obtain the asymptotic loss of $2V_{\mathcal{C}}$ for the mixture predictor. While not a problem for the case $V_{\mathcal{C}} = 0$, this is of course not what we want in the general case. The reason behind this problem is that for the construction in that proof one can only use the fact that each of the measures μ_k in the sequence is as good as the predictor ρ whose existence is assumed (the one that attains $V_{\mathcal{C}} = 0$). In contrast, in the proof below we are able to use the fact that each measure in the sequence is in fact much better than ρ on some subsets of $\mathcal{X}^n$.

Proof Define the weights $w_k := w/(k+1)\log^2(k+1)$, where w is the normalizer such that $\sum_{k\in\mathbb{N}} w_k = 1$. Introduce the notation $M := \log|\mathcal{X}|$.

We start with the second statement of the theorem. Take any predictor ρ. We shall find a measure ν of the form $\sum_{k\in\mathbb{N}} w'_k\mu_k$, where $\mu_k \in \mathcal{C}$ such that

$$D(\mu,\nu) \le D(\mu,\rho)\ \forall \mu \in \mathcal{C}. \tag{4}$$

Replacing ρ with $(\rho+p)/2$ if necessary, where p is the i.i.d. measure with equal probabilities of outcomes, we shall assume

$$-\log\rho(x_{1..n}) \le nM+1 \text{ for all } n\in\mathbb{N} \text{ and } x_{1..n}\in\mathcal{X}^n. \tag{5}$$

Thus, in particular, $d_n(\mu,\rho) \le nM+1$ for all μ.

For each $\mu\in\mathcal{C}$, $n\in\mathbb{N}$ define the sets

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

From the Markov's inequality, we obtain

$$\mu(\mathcal{X}^n\backslash T^n_\mu) \le 1/n. \tag{7}$$

For each $k\in\mathbb{N}$ let U_k be the partition of $[-\frac{\log n}{n}, M+\frac{1}{n}]$ into k intervals defined as follows. $U_k := \{u^i_k : i = 1..k\}$, where

$$u^i_k = \begin{cases} \left[-\frac{\log n}{n}, \frac{iM}{k}\right] & i=1,\\ \left(\frac{(i-1)M}{k}, \frac{iM}{k}\right] & 1<i<k,\\ \left(\frac{(i-1)M}{k}, M+\frac{1}{n}\right] & i=k. \end{cases}$$

Thus, U_k is a partition of $[0,M]$ into k equal intervals but for some padding that we added to the leftmost and the rightmost intervals: on the left we added $[-\frac{\log n}{n}, 0)$ and on the right $(M, M+1/n]$.

For each $\mu\in\mathcal{C}$, $n,k\in\mathbb{N}$, $i=1..k$ define the sets

$$T^n_{\mu,k,i} := \left\{x_{1..n}\in\mathcal{X}^n : \frac{1}{n}\log\frac{\mu(x_{1..n})}{\rho(x_{1..n})} \in u^i_k\right\}. \tag{8}$$

Observe that, for every $\mu, k, n \in \mathbb{N}$, these sets constitute a partition of T^n_μ into k disjoint sets: indeed, on the left we have $\frac{1}{n}\log\frac{\mu(x_{1..n})}{\rho(x_{1..n})} \geq -\frac{1}{n}\log n$ by definition (6) of T^n_μ, and on the right we have $\frac{1}{n}\log\frac{\mu(x_{1..n})}{\rho(x_{1..n})} \leq M + 1/n$ from (5). In particular, from this definition, for all $x_{1..n} \in T^n_{\mu,k,i}$ we have

$$\mu(x_{1..n}) \leq 2^{\frac{iM}{k}n+1}\rho(x_{1..n}). \tag{9}$$

For every $n, k \in \mathbb{N}$ and $i \in \{1..k\}$ consider the following construction. Define $m_1 := \max_{\mu\in\mathcal{C}} \rho(T^n_{\mu,k,i})$ (since the sets $\mathcal{X}^n$ are finite all suprema are reached). Find any μ_1 such that $\rho(T^n_{\mu_1,k,i}) = m_1$ and let $T_1 := T^n_{\mu_1,k,i}$. For $l > 1$, let $m_l := \max_{\mu\in\mathcal{C}} \rho(T^n_{\mu,k,i}\backslash T_{l-1})$. If $m_l > 0$, let μ_l be any $\mu \in \mathcal{C}$ such that $\rho(T^n_{\mu_l,k,i}\backslash T_{l-1}) = m_l$, and let $T_l := T_{l-1} \cup T^n_{\mu_l,k,i}$; otherwise let $T_l := T_{l-1}$ and $\mu_l := \mu_{l-1}$. Note that, for each $x_{1..n} \in T_l$ there is $l' \leq l$ such that $x_{1..n} \in T^n_{\mu_{l'},k,i}$ and thus from (8) we get

$$2^{\frac{(i-1)M}{k}n-\log n}\rho(x_{1..n}) \leq \mu_{l'}(x_{1..n}). \tag{10}$$

Finally, define

$$\nu_{n,k,i} := \sum_{l=1}^{\infty} w_l\mu_l \tag{11}$$

(notice that for every n, k, i there is only a finite number of positive m_l, since the set $\mathcal{X}^n$ is finite; thus the sum in the last definition is effectively finite) and the resulting predictor ν as

$$\nu := \frac{1}{2}\sum_{n,k\in\mathbb{N}} w_n w_k \frac{1}{k}\sum_{i=1}^{k}\nu_{n,k,i} + \frac{1}{2}r, \tag{12}$$

where r is a regularizer to be defined below. The regularizer r is defined so as to have for each $\mu' \in \mathcal{C}$ and $n \in \mathbb{N}$

$$\log\frac{\mu'(x_{1..n})}{\nu(x_{1..n})} \leq nM + O(\log n) \quad \text{for all } x_{1..n} \in \mathcal{X}^n; \tag{13}$$

this and the stronger statement (5) for ν can be obtained analogously to the latter inequality in the case the i.i.d. measure p is in $\mathcal{C}$; otherwise (since we need to define ν as a combination of measures from $\mathcal{C}$ only), r can be defined the same way as is done in (Ryabko, 2010, Step r of the proof of Theorem 5); for the sake of completeness, this argument is given in the end of this proof.

We shall show that (4) holds for ν defined in (12), which will establish the second statement of the theorem. First, we want to show that, for each $\mu \in \mathcal{C}$, for each fixed k, i, the sets $T^n_{\mu,k,i}$ are covered up to a negligible μ-probability by the sets T_l with indices l that are not too small. Observe that, by definition, for each n, i, k the sets $T_l\backslash T_{l-1}$ are disjoint (for different l) and have non-increasing (with l) ρ-probability. Therefore, $\rho(T_{l+1}\backslash T_l) \leq 1/l$ for all $l \in \mathbb{N}$. Hence, from the definition of T_l, we must also have $\rho(T^n_{\mu,k,i}\backslash T_l) \leq 1/l$ for all $l \in \mathbb{N}$. From the latter inequality and (9) we obtain $\mu(T^n_{\mu,k,i}\backslash T_l) \leq \frac{1}{l}2^{\frac{iM}{k}n+1}$. Consequently, for any $a > M/k$ taking $l := 2^{(\frac{iM}{k}+a)n+1}$ we obtain that for each $x_{1..n} \in T^n_{\mu,k,i}$ except

possibly for a set of μ-probability 2^{-an} (that is, for $x_{1..n} \in T^n_{\mu,k,i}\backslash T_l$) there is $l' \le l$ such that the following chain holds

$$\nu(x_{1..n}) \ge w_n w_k \frac{1}{k} w_{2^{(\frac{iM}{k}+a)n}} \mu_{l'}(x_{1..n}) \ge 2^{-(\frac{iM}{k}+a)n+o(n)} \mu_{l'}(x_{1..n}) \ge 2^{-(a+\frac{M}{k})n+o(n)} \rho(x_{1..n}), \tag{14}$$

where the first inequality is from (12) and (11) (with the value of l we selected), the second is by definition of w_l and the third uses (10).

Suppose that there exist $\mu \in \mathcal{C}$ and $\delta'' > 0$ such that

$$\frac{1}{n} d_n(\mu, \nu) > \frac{1}{n} d_n(\mu, \rho) + \delta'' \tag{15}$$

infinitely often. Note that (13) implies that $\frac{1}{n} \log \frac{\mu(x_{1..n})}{\nu(x_{1..n})}$ is bounded uniformly (w.r.t. $x_{1..n}$) from above. Using this and (7) we can restrict (15) to T^n_μ, obtaining

$$\frac{1}{n} \sum_{x_{1..n} \in T^n_\mu} \mu(x_{1..n}) \log \frac{\mu(x_{1..n})}{\nu(x_{1..n})} > \frac{1}{n} \sum_{x_{1..n} \in T^n_\mu} \mu(x_{1..n}) \log \frac{\mu(x_{1..n})}{\rho(x_{1..n})} + \delta' \tag{16}$$

infinitely often for some $\delta' > 0$. Consequently, again invoking the boundedness (13), we conclude that there exist $\varepsilon', \delta > 0$, an infinite sequence of indices $(n'_j)_{j\in\mathbb{N}}$ and sets $A'_j \subset T^{n'_j}_\mu$ such that $\mu(A'_j) > \varepsilon'$ and $-\log \nu(x_{1..n_j}) > n_j\delta - \log \rho(x_{1..n_j})$ for $x_{1..n_j} \in A'_j$. Recall that for each $k \in \mathbb{N}$ the sets $T^n_{\mu,k,j}$ partition each of the sets T^n_μ and therefore each of the sets A'_j into at most k sets. Hence, for every k there must exist a cell of this partition, that is, an index $i \in \{1..k\}$, along with an $\varepsilon \ge \varepsilon'/k > 0$ and subsequences $(n_j)_{j\in\mathbb{N}}$ and $(A_j)_{j\in\mathbb{N}}$ (with $A_j \subset X^{n_j}$) of the sequences $(n'_j)_{j\in\mathbb{N}}$ and $(A'_j)_{j\in\mathbb{N}}$ such that $\mu(A_j \cap T^{n_j}_{\mu,k,i}) > \varepsilon$ for all $j \in \mathbb{N}$. Denote $B_j = A_j \cap T^{n_j}_{\mu,k,i}$ for each $j \in \mathbb{N}$. We have thus obtained, finally, an infinite sequence of indices $(n_j)_{j\in\mathbb{N}}$ and sets $B_j \subset T^{n_j}_{\mu,k,i}$ of μ-probability bounded from below by $\varepsilon/2$, such that for each $x_{1..n_j} \in B_j$ we have

$$\nu(x_{1..n_j}) < 2^{-\delta n_j} \rho(x_{1..n_j}). \tag{17}$$

Take $k > 0$ such that $M/k < \delta/4$. To conclude the proof of the second statement of the theorem, it remains to observe that (17) contradicts (14) with $a = \delta/2$.

To prove the first statement of the theorem, first, let $\gamma_j > V_\mathcal{C}$, $j \in \mathbb{N}$ be a sequence such that $\lim_{j\to\infty} \gamma_j = V_\mathcal{C}$. Find then a sequence $\rho_j \in \mathcal{P}$ such that $D(\mathcal{C}, \rho_j) \le \gamma_j$. We need to find a predictor φ of the form $\sum_{k\in\mathbb{N}} w'_k \mu_k$, where $\mu_k \in \mathcal{C}$ such that $D(\mathcal{C}, \varphi) \le D(\mathcal{C}, \rho_j)$ for every $j \in \mathbb{N}$. So far we have shown that for every ρ_j, $j \in \mathbb{N}$ there is a measure ν_j of the required form such that $D(\mathcal{C}, \nu_j) \le D(\mathcal{C}, \rho_j)$. It remains to define $\varphi := \sum_{j\in\mathbb{N}} w_j \nu_j$. Indeed, for every $\mu \in \mathcal{C}$ and every $j \in \mathbb{N}$ we have

$$d_n(\mu, \varphi) = E_\mu \log \frac{\mu(x_{1..n})}{\varphi(x_{1..n})} \le E_\mu \log \frac{\mu(x_{1..n})}{\nu_j(x_{1..n})} - \log w_j,$$

so that $D(\mu, \varphi) \le D(\mu, \nu_j) \le D(\mu, \rho_j) \le \gamma_j$. Finally, recall that $\gamma_j \to V_\mathcal{C}$ to obtain the desired statement.

It remains to come back to (13) and define the regularizer r as a combination of measures from $\mathcal{C}$ for this inequality to hold. For each $n \in \mathbb{N}$, denote

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

and let for each $x_{1..n} \in \mathcal{X}^n$ the measure $\mu_{x_{1..n}}$ be any measure from $\mathcal{C}$ such that $\mu_{x_{1..n}}(x_{1..n}) \geq \frac{1}{2}\sup_{\mu\in\mathcal{C}} \mu(x_{1..n})$. Define

$$r'_n := \frac{1}{|A_n|}\sum_{x_{1..n}\in A_n} \mu_{x_{1..n}}$$

for each $n \in \mathbb{N}$, and let $r := \sum_{n\in\mathbb{N}} w_n r'_n$. For every $\mu \in \mathcal{C}$ we have

$$r(x_{1..n}) \geq w_n |A_n|^{-1} \mu_{x_{1..n}}(x_{1..n}) \geq \frac{1}{2} w_n |\mathcal{X}|^{-n} \mu(x_{1..n})$$

for every $n \in \mathbb{N}$ and every $x_{1..n} \in A_n$, establishing (13). ■

4. Decision-theoretic interpretations

Classical decision theory is concerned with single-step games. Among its key results are the complete class and minimax theorems. The infinite-horizon case studied here presents both differences and similarities which we attempt to summarize here. A distinction worth mentioning at this point is that the results presented here are obtained under no assumptions whatsoever, whereas the results in decision theory we refer to always have a number of conditions; on the other hand, here we are concerned with just one specific loss function (KL divergence) rather than general losses as is common in decision theory.

Predictors $\rho \in \mathcal{P}$ are called *strategies of the statistician*. The measures $\mu \in \mathcal{C}$ are now the basic *strategies of the opponent*, and the first thing we need to do is to extend these to randomized strategies. To this end, denote $\mathcal{C}^*$ the set of all probability distributions over measurable subsets of $\mathcal{C}$. Thus, the opponent selects a randomized strategy $W \in \mathcal{C}^*$ and the statistician (predictor) ρ suffers the loss

$$E_{W(\mu)} D(\mu, \rho), \tag{18}$$

where the notation $W(\mu)$ means that μ is drawn according to W. Note a distinction with the combinations we considered before. A combination of the kind $\nu = \int_{\mathcal{C}} dW$ is itself a probability measure over the one-way infinite sequences, whereas a measure $W \in \mathcal{C}^*$ is a measure over $\mathcal{C}$. In other words, the difference is between putting the integral $\int_{\mathcal{C}} dW$ outside of D as in (18) or inside of D which would be $D(\int_{\mathcal{C}} dW(\mu), \rho)$. In the terminology of Gray (1988), the measure $\int_{\mathcal{C}} dW(\mu) \in \mathcal{P}$ is the barycentre of $W \in \mathcal{C}^*$.

Minimax. Generalizing the definition (3) of $V_{\mathcal{C}}$, we can now introduce the *upper value*

$$\bar{V}_{\mathcal{C}} := \inf_{\rho\in\mathcal{P}} \sup_{W\in\mathcal{C}^*} E_{W(\mu)} D(\mu, \rho). \tag{19}$$

Furthermore, the *maximin* (the *lower value*) is defined as

$$\underline{V}_{\mathcal{C}} := \sup_{W\in\mathcal{C}^*} \inf_{\rho\in\mathcal{P}} E_{W(\mu)} D(\mu, \rho). \tag{20}$$

The so-called minimax theorems in decision theory (e.g., Ferguson, 1967) for single-step games and general loss functions state that, under certain conditions, $\bar{V}_{\mathcal{C}} = \underline{V}_{\mathcal{C}}$ and the statistician has a minimax strategy, that is, there exists ρ on which $\bar{V}_{\mathcal{C}}$ is attained. Minimax theorems generalize the classical result of von Neumann (1928), and provide sufficient conditions of various generality for it to hold. A rather general sufficient condition is the existence of a topology with respect to which the set of all strategies of the statistician, $\mathcal{P}$ in our case, is compact, and the risk, which is $D(\mu, \rho)$ in our case, is lower semicontinuous. Such a condition seems nontrivial to verify. For example, a (meaningful) topology with respect to which $\mathcal{P}$ is compact is that of the so-called distributional distance (Gray, 1988) (in our case it coincides with the topology of the weak* convergence), but $D(\mu, \rho)$ is not (lower) semicontinuous with respect to it. Some other (including non-topological) sufficient conditions are given in (Sion, 1958; LeCam, 1955).

In our setup, it is easy to see that $\bar{V}_{\mathcal{C}} = V_{\mathcal{C}}$ and so Theorem 3 holds for $\bar{V}_{\mathcal{C}}$. Thus, using decision-theoretic terminology, we can state the following.

Corollary 4 (minimax) *For every set $\mathcal{C}$ of strategies of the opponent, the statistician has a minimax strategy.*

However, the question of whether the upper and the lower values coincide remains open. That is, we are taking the worst possible distribution over $\mathcal{C}$, and ask what is the best possible predictor with the knowledge of this distribution ahead of time. The question is whether $\underline{V}_{\mathcal{C}} = V_{\mathcal{C}}$. A closely related question is whether there is a worst possible strategy for the opponent. This latter would be somehow a maximally spread-out (or maximal entropy) distribution over $\mathcal{C}$. In general, measurability issues seem to be very relevant here, especially for the maximal-entropy question.

Complete class. For a set of measures (strategies of the opponent) $\mathcal{C}$, a predictor ρ_1 is said to be *as good as* a predictor ρ_2 if $D(\mu, \rho_1) \le D(\mu, \rho_2)$ for all $\mu \in \mathcal{C}$. A predictor ρ_1 is *better (dominates)* ρ_2 if ρ_1 is as good as ρ_2 and $D(\mu, \rho_1) < D(\mu, \rho_2)$ for some $\mu \in \mathcal{C}$. A predictor ρ is *admissible* (also called *Pareto optimal*) if there is no predictor ρ' which is better than ρ; otherwise it is called *inadmissible*. Similarly, a set of predictors D is called a *complete class* if for every $\rho' \notin D$ there is $\rho \in D$ such that ρ is better than ρ'. A set of of predictors D is called an *essentially complete class* if for every $\rho' \notin D$ there is $\rho \in D$ such that ρ is as good as ρ'. An (essentially) complete class is called *minimal* if none of its proper subsets is (essentially) complete.

Furthermore, in decision-theoretic terminology (Ferguson, 1967), a predictor ρ is called a *Bayes rule* for a prior $W \in \mathcal{C}^*$ if it is optimal for W, that is, if it attains $\inf_{\rho \in \mathcal{P}} E_{W(\mu)} D(\mu, \rho)$. Clearly, if W is concentrated on a finite or countable set then any mixture over this set with full support is a Bayes rule, and the value of the inf above is 0.

In decision theory, the complete class theorem (Wald, 1950; LeCam, 1955, see also Ferguson, 1967) states that, under certain conditions similar to those above for the minimax theorem, the set of Bayes rules is complete and the admissible Bayes rules form a minimal complete class.

An important difference in our set-up is that all strategies are inadmissible (unless $V_{\mathcal{C}} = 0$), and one cannot speak about minimal (essentially) complete classes. However, the set of all Bayes rules is still essentially complete, and an even stronger statement holds: it is enough to consider all Bayes rules with countable priors:

Proposition 5 *For every set $\mathcal{C}$, the set of those Bayes rules whose priors are concentrated on at most countable sets is essentially complete. There is no admissible rule (predictor) and no minimal essentially complete class unless $V_{\mathcal{C}} = 0$. In the latter case, every predictor ρ that attains this value is admissible and the set $\{\rho\}$ is minimal essentially complete.*

Proof The first statement is a reformulation of the second statement of Theorem 3. To prove the second statement, consider any $\mathcal{C}$ such that $V_{\mathcal{C}} > 0$, take a predictor ρ that attains this value (such a predictor exists by Theorem 3), and a measure μ such that $D(\mu, \rho) > 0$. Then for a predictor $\rho' := (\rho + \mu)/2$ we have $D(\mu, \rho') = 0$ while $D(\nu, \rho) \geq D(\nu, \rho')$ for all ν, so that ρ' is better than ρ. Therefore, ρ is inadmissible. The statement about minimal essentially complete class is proven analogously. The statement about the case $V_{\mathcal{C}} = 0$ follows directly from the definitions. ■

5. Examples

Ryabko (2010) considers in detail several examples of processes classes for the case $V_{\mathcal{C}} = 0$; these examples include countable sets $\mathcal{C}$, the set of i.i.d. measures, Markov chains, bounded-memory processes and stationary ergodic processes. Therefore, here we will only look at the case $V_{\mathcal{C}} > 0$. For simplicity, we assume $\mathcal{X} = \{0, 1\}$ in the examples.

Typical Bernoulli 1/3 sequences. We start with a somewhat artificial example, but a one on which it is relatively easy to see how countable mixtures give predictors for large uncountable sets. Take the binary $\mathcal{X}$ and consider all sequences $\mathbf{x} \in \mathcal{X}^\infty$ such that the limiting number of 1s in $\mathbf{x}$ equals 1/3. Denote the set of these sequences S and let the set $\mathcal{C}$ consist of all Dirac measures concentrated on sequences from S. Observe that the Bernoulli i.i.d. measure $\delta_{1/3}$ with probability 1/3 of 1 predicts measures in $\mathcal{C}$ relatively well: $D(\mathcal{C}, \delta_{1/3}) = h(1/3)$, where h stands for the binary entropy, and this is also the minimax loss for this set, $V_{\mathcal{C}}$. It might then appear surprising that this loss is achievable by a combination of countably many measures from $\mathcal{C}$ — after all, this set consists only of deterministic measures. Let us try to see what such a combination may look like. By definition, for any sequence $\mathbf{x} \in S$ and every ε we can find $n_\varepsilon(\mathbf{x}) \in \mathbb{N}$ such that for all $n \geq n_\varepsilon(\mathbf{x})$ the average number of 1s in $x_{1..n}$ is within ε of 1/3. Fix the sequence of indices $k_j := 2^j$, $j \in \mathbb{N}$ and the sequence of thresholds $\varepsilon_l := 2^{-l}$. For each k_j let ${S'}_j^l \subset S$ be the set of all sequences $\mathbf{x} \in S$ such that $n_{\varepsilon_l}(\mathbf{x}) < n_j$. Select then a finite subset S_j^l of ${S'}_j^l$ such that for each $\mathbf{x}' \in {S'}_j^l$ there is $\mathbf{x} \in S$ such that $x'_{1..n_j} = x_{1..n_j}$. This is of course possible since the set $\mathcal{X}^{n_j}$ is finite. Now for each $\mathbf{x} \in S_j^l$ take the corresponding measure $\mu_{\mathbf{x}} \in \mathcal{C}$ and attach to it the weight $w_l w_j / |S_j^l|$, where, as before, we are using the weights $w_k = w/k \log^2 k$. Taking these measures for all $j, l \in \mathbb{N}$ we obtain our convex combination. Of course we did not enumerate all sequences in S (or measures in $\mathcal{C}$) this way; but for each sequence $\mathbf{x} \in S$ and for each n there is a sequence among those that we did enumerate that coincides with $\mathbf{x}$ up to the index n. One can then use the theory of types (Csiszar, 1998) to calculate the sizes of the sets S_j^l and to check that the weights we found give the optimal loss we are after; but for the illustrative purposes of this example this is already not necessary.

Processes with abrupt changes. Start with a family of distributions S for which we have a good predictor, for example S is the set B of all Bernoulli i.i.d. processes, or more

generally a set for which $V_S = 0$. The family $\mathcal{C}_{S,\alpha}$ parametrized by $\alpha \in (0,1)$ and S is then the family of all processes constructed as follows: there is a sequence of indexes n_i such that $X_{n_i..n_{i+1}}$ is distributed according to μ_i for some $\mu_i \in S$. Take then all possible sequences μ_i and all sequences n_i whose limiting frequency $\lim_{i\to\infty}\{i : n_i < n\}$ is bounded by α to obtain our set $\mathcal{C}_{S,\alpha}$. Thus, we have a family of processes with abrupt changes in distribution, where between the changes the distribution is from S, the changes are assumed to have the frequency bounded by α but are otherwise arbitrary. This example was considered in (Willems, 1996) for the case $S = B$, with the goal of minimizing the regret w.r.t. the predictor that knows where the changes occur (the value $V_{\mathcal{C}}$ was not considered directly). The method proposed in the latter work, in fact, is not limited to the case $S = B$, but is general. The algorithm is based on a prior over all possible sequences n_i of changes; between the changes the optimal predictor for B is used, which is also a Bayesian predictor with a specific prior. The regret obtained is of order $\log n$. Since for Bernoulli processes themselves the best achievable loss up to time n is $1/2 \log n + 1$, we can see that $V_{\mathcal{C}_B,\alpha} = \alpha(1 - 1/2 \log \alpha)$. A similar result can be obtained if we replace Bernoulli processes with Markov processes, but not with an arbitrary S for which $V_S = 0$. For example, if we take S to be all finite-memory distributions, then the resulting process may be completely unpredictable ($V_{\mathcal{C}} = 1$): indeed, if the memory of distributions μ_i grows (with i) faster than αn, then there is little one can do. For such sets S one can make the problem amenable by restricting the way the distributions μ_i are selected, for example, imposing an ergodicity-like condition that the average distribution has a limit. Another way (often considered in the literature in slightly different settings, see (Gyorgy et al., 2012) and references therein) is to have $\alpha \to 0$, although in this case one recovers $V_{\mathcal{C}_S} = 0$ provided α goes to 0 slowly enough (and, of course, provided $V_S = 0$).

Predictable aspects. The preceding example can be thought of as an instantiation of the general class of processes in which some aspects are predictable while others are not. Thus, in the considered example changes between the distributions were unpredictable, but between the changes the distributions were predictable. Another example of this kind is that of processes predictable on some scales but not on others. Imagine that it is possible to predict, for example, large fluctuations of the process but not small fluctuations (or the other way around). More formally, consider now an alphabet $\mathcal{X}$ with $|\mathcal{X}| > 2$, and let Y be some partition of $\mathcal{X}$. For any sequence $x_1, \dots, x_n, \dots$ there is an associated sequence $y_1, \dots, y_n, \dots$ where y_i is defined as $y \in Y$ such that $x_i \in y$. We can obtain examples of sets $\mathcal{C}$ of processes with $V_{\mathcal{C}} \in (0,1)$ by restricting the distribution of $y_1, \dots, y_n, \dots$ to a set B with $V_B = 0$. The interpretation is that, again, we can model the y part (by processes in B) but not the rest, which we then allow to be arbitrary.

Yet another example is that of processes predictable only after certain kind of events: for example, after a price drop; or after a rain. At other times, the process is unpredictable: it can, again, be an arbitrary deterministic sequence. More formally, let a set $A \subset \mathcal{X}^* := \cup_{k\in\mathbb{N}}\mathcal{X}^k$ be measurable. Consider for each sequence $\mathbf{x} = x_1, \dots, x_n, \dots$ another (possibly finite) sequence $\mathbf{x}' = x'_1, \dots, x'_n, \dots$ given by $x'_i := (x_{n_i+1})_{i\in\mathbb{N}}$ where n_i are all indexes such that $x_{1..n_i} \in A$. We now form the set $\mathcal{C}$ as the set of all processes μ such that $\mathbf{x}'$ belongs (μ-a.s.) to some pre-defined set B; for this set B we may have $V_B = 0$. This means that we can model what happens after events in A — by processes in B, but not the rest of the times, on which we say the process may be arbitrary. For different A and B we then obtain examples

where $V_{\mathcal{C}} \in (0,1)$. A particular instance of this, as well as of the preceding, example is predicting selected bits. Specifically, Hutter (2005) mentions the problem of predicting every second bit in a sequence, asking whether Solomonoff's predictor (Solomonoff, 1978), which is a mixture over the countable set of all (semi-)computable measures, would still work in the setting — despite the sequence being possibly non-computable. (For the time-averaged KL loss of the present work this question is trivial, but is not trivial for other losses.) This question is generalized (and answered) to that of predicting the bits on a computable subsequence of time steps by Lattimore et al. (2011).

6. Relation to the non-realizable case

As mentioned in the Introduction, Ryabko (2016) shows that in the non-realizable case all Bayesian mixture predictors may be suboptimal. Here we make this statement precise in order to clarify its relation to the main result.

The non-realizable case is when the measure generating the data does not belong to $\mathcal{C}$. We are then looking at the set $\mathcal{C}$ as the set of experts or models, and we seek a predictor ρ that predicts any measure ν (that generates the data) whatsoever as well as the best (for this ν) $\mu \in \mathcal{C}$.

Thus, if we have two predictors μ and ρ, we can define the *regret* up to time n of (using the predictor) ρ as opposed to (using the predictor) μ on the measure ν (that is, ν generates the sequence to predict) as

$$R_n^\nu(\mu,\rho) := d_n(\nu,\rho) - d_n(\nu,\mu).$$

Furthermore, define the asymptotic average regret as

$$\bar{R}^\nu(\mu,\rho) := \limsup_{n\to\infty} \frac{1}{n} R_n^\nu(\mu,\rho),$$

and

$$\bar{R}^\nu(C,\rho) := \sup_{\mu\in C} \bar{R}^\nu(\mu,\rho).$$

Ryabko (2016) shows that there exists a set $\mathcal{C}$ such that any Bayesian predictor must have a linear regret, while there exists a predictor with a sublinear regret:

Theorem 6 (Ryabko, 2016) *There exist a set C of measures and a predictor ρ such that $\bar{R}^\nu(C,\rho) = 0$ for every measure ν, yet for every Bayesian predictor φ with a prior concentrated on C there exists a measure ν such that we have $\bar{R}^\nu(C,\varphi) \geq c > 0$ where c is a constant (independent of φ).*

In the same work it is argued that this applies more broadly than just to Bayesian predictors: all non-trivial combinations of measures in $\mathcal{C}$ may be useless for minimizing regret. We remind again that such a set $\mathcal{C}$ must necessarily be uncountable.

7. Discussion and directions for future research

A statistician facing an unknown stochastic phenomenon has a large, nonparametric model class at hand that she has reasons to believe captures some aspects of the problem. Yet other

aspects remain completely enigmatic, and there is little hope that the process generating the data indeed comes from the model class. For this reason the statistician is content at having non-zero error no matter how much data may become available now or in the future, but she would still like to make some use of the model. There are now two rather distinct ways to proceed. One is to say that the data may come from an arbitrary deterministic sequence, and try to construct a predictor that minimizes the regret with respect to every distribution in the model class, on every deterministic sequence. The other way is to try to enlarge the model class, in particular, by allowing that all there is unknown in the process may be arbitrary (that is, an arbitrary deterministic sequence). This second way may be more difficult precisely on the modelling step. Yet, the conclusion of this work is that this is the way to follow, for in this case one can be sure that it is possible to make statistical inference by standard available tools, specifically, Bayesian forecasting: even if the best achievable asymptotic error is non-zero it is attained by a Bayesian forecaster with some prior. Finding such a prior is a separate problem, but it is a one with which Bayesians are familiar. Here, modelling that unknown part should not create much trouble: a good distribution over all deterministic sequences is just the Bernoulli i.i.d. measure with equiprobable outcomes. (Note that it is not necessary to look for priors concentrated on countable sets.) On the other hand, for the regret-minimization route, the statistician cannot use an arbitrary model class; indeed, she would first need to make sure that regret minimization is viable at all for the model class at hand: it may happen that every combination of distributions in the model is suboptimal.

There are no criteria for checking this, only some (rather small) examples, such as finite or countable sets, or specific parametric families. Finding such criteria for the viability of regret minimization is an interesting open problem. To make it more precise, the question is for which sets $\mathcal{C}$ of distributions the minimax regret (is attainable and) can be attained by a combination (either Bayesian or some other) of distributions in $\mathcal{C}$.

It is worth noting that the conclusions of the paper are not about Bayesian versus non-Bayesian inference. Rather, Bayesian inference is used as a generic approach to construct predictors for general (uncountable) model classes. At this level of generality it is hard to find any alternative approach, although it would be interesting to see which predictors can be generalized (to arbitrary model classes) and whether the corresponding result holds for them. The negative result of Ryabko (2016), as explained in that work, is not restricted to Bayesian predictors but holds much more generality.

The quantity $V_{\mathcal{C}}$ introduced in Section 3 appears to be an important characteristic of a set of processes $\mathcal{C}$, as it quantifies the best possible attainable worst-case performance. It thus seems interesting to calculate it for various classes of processes, similar to what is done in Section 5 for the examples considered there.

Another interesting question concerns different losses. While the proof does not seem to be hinged very specifically on the log loss, it does use some properties of it in an important way. In particular, the property that if μ predicts ν then also any convex combination $\alpha\mu + (1-\alpha)\rho$ predicts ν for any ρ. This does not hold for some other losses, in particular already for KL loss without Cesaro averaging; see (Ryabko and Hutter, 2008) for some results on this property.

Finally, an intriguing question is whether a result like Theorem 3 holds if one allows convergence rates into consideration. Now that we know that the minimax asymptotic

error is achievable, we can ask whether the minimax rate of convergence to this error is also achievable (by a Bayesian predictor). The proof of the version of Theorem 3 for the $V_{\mathcal{C}} = 0$ case in (Ryabko, 2010) clearly does not generalize to achieve such a result (the rates one extract from that proof are rather bad), but with the present proof this may be possible.

References

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

I. Csiszar. The method of types. *IEEE Transactions on Information Theory*, 44(6):2505–2523, 1998.

Thomas S Ferguson. *Mathematical statistics: A decision theoretic approach*. Academic press, 1967.

R. G. Gallager. Source coding with side information and universal coding. Technical Report LIDS-P-937, M.I.T., 1976 (revised 1979).

Robert M. Gray. *Probability, Random Processes, and Ergodic Properties*. Springer Verlag, 1988.

P. Grünwald and Ph. Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Annals of Statistics*, pages 1367–1433, 2004.

L. Gyorfi and G. Ottucsak. Sequential prediction of unbounded stationary time series. *Information Theory, IEEE Transactions on*, 53(5):1866 –1872, May 2007. ISSN 0018-9448. doi: 10.1109/TIT.2007.894660.

András Gyorgy, Tamás Linder, and Gábor Lugosi. Efficient tracking of large classes of experts. *IEEE Transactions on Information Theory*, 58(11):6709–6725, 2012.

D. Haussler. A general minimax result for relative entropy. *IEEE Trans. on Information Theory*, 43(4):1276–1280, 1997.

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

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

Tor Lattimore, Marcus Hutter, and Vaibhav Gavane. Universal prediction of selected bits. In *Algorithmic Learning Theory*, pages 262–276. Springer, 2011.

Lucien LeCam. An extension of Wald's theory of statistical decision functions. *The Annals of Mathematical Statistics*, 26(1):69–81, 1955.

G. Morvai, S.J. Yakowitz, and P. Algoet. Weakly convergent nonparametric forecasting of stationary time series. *Information Theory, IEEE Transactions on*, 43(2):483 –498, March 1997. ISSN 0018-9448. doi: 10.1109/18.556107.

J v Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928.

Yuichi Noguchi. Merging with a set of probability measures: A characterization. *Theoretical Economics*, 10(2):411–444, 2015.

Boris Ryabko. Coding of a source with unknown but ordered probabilities. *Problems of Information Transmission*, 15(2):134–138, 1979.

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

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

Daniil Ryabko. On the relation between realizable and non-realizable cases of the sequence prediction problem. *Journal of Machine Learning Research*, 12:2161–2180, 2011.

Daniil Ryabko. Things Bayes can't do. In *Proceedings of the 27th International Conference on Algorithmic Learning Theory (ALT'16)*, volume 9925 of *LNCS*, pages 253–260, Bari, Italy, 2016. Springer.

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

Maurice Sion. On general minimax theorems. *Pacific J. Math*, 8(1):171–176, 1958.

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

Abraham Wald. *Statistical decision functions*. John Wiley&Sons, New York, 1950.

Frans MJ Willems. Coding for a binary independent piecewise-identically-distributed source. *IEEE Transactions on Information Theory*, 42(6):2210–2217, 1996.