

Finite-time optimality of Bayesian predictors

Daniil Ryabko

Abstract

The problem of sequential probability forecasting is considered in the most general setting: a model set \mathcal{C} is given, and it is required to predict as well as possible if any of the measures (environments) in \mathcal{C} is chosen to generate the data. No assumptions whatsoever are made on the model class \mathcal{C} , in particular, no independence or mixing assumptions; \mathcal{C} may not be measurable; there may be no predictor whose loss is sublinear, etc. It is shown that the cumulative loss of any possible predictor can be matched by that of a Bayesian predictor whose prior is discrete and is concentrated on \mathcal{C} , up to an additive term of order $\log n$, where n is the time step. The bound holds for every n and every measure in \mathcal{C} . This is the first non-asymptotic result of this kind. In addition, a non-matching lower bound is established: it goes to infinity with n but may do so arbitrarily slow.

1 Introduction

Choosing a model is a hard problem. Its solutions are often driven by the ease of finding an algorithm rather than by the adequacy of the model to the task at hand. In the context of the prediction problem, at the very least it is typically assumed that a predictor whose loss is sublinear exists. Even under this assumption, there are no generic methods for constructing a predictor given only a model set \mathcal{C} . This applies not only to the prediction problem, but more generally. One generic method for constructing a learning algorithm is the Bayesian one: choose a prior over the model class and predict according to the posterior distribution given the data. However, there are classical results that show that some, or even, in some sense, most of the priors result in an inconsistent method (Freedman, 1963, 1965; Diaconis und Freedman, 1986). Therefore, the question arises: is it possible to show that the Bayesian with at least *some* prior will be optimal? In the asymptotic sense, this question was answered in the positive by Ryabko (2010, 2017) (first under the assumption that the best achievable asymptotic average error is 0, and then without this assumption). Thus, the smallest asymptotic average error is achieved by a Bayesian predictor with some prior. However, this leaves open the question of what happens before infinity, allowing for the possibility that, for finite n , every Bayesian predictor is grossly suboptimal.

Here we resolve this doubt, and show that, for any model set \mathcal{C} , there is a prior over this set, such that the Bayesian predictor with this prior has optimal

cumulative error up to an additive term of order $\log n$ for every time step n for every measure μ in \mathcal{C} (not just with prior average). This means that, the regret of being Bayesian, with some prior, on time-average is at most $O(\log n/n)$. This is generally considered rather small; in particular, already the fact that there is no multiplicative factor may be remarkable, and already this is new for the case of $o(n)$ cumulative loss. We also establish a lower bound, though the gap with respect to the upper bound is relatively large: the lower bound on the cumulative regret of being a Bayesian goes to infinity but may do so arbitrarily slow.

Setup. A bit more formally, the problem is that of sequential probability forecasting in the following setting. Given a sequence x_1, \dots, x_n of observations $x_i \in \mathcal{X}$, where \mathcal{X} is a finite set, it is required 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, $n = 1, 2, \dots$. The problem is considered in full generality; in particular, outcomes may exhibit arbitrary dependence. What is given is a set \mathcal{C} of measures over the space of all one-way infinite sequences. It is assumed that one of the measures in \mathcal{C} , say μ , is chosen to generate the data, and it is required to construct a predictor whose error is as small as possible for every $\mu \in \mathcal{C}$. The error is measured in terms of the expected (with respect to the unknown μ measure that generates the data) cumulative (over n time steps) KL divergence (log loss) L_n :

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

Motivation. This and related problems arise in a variety of applications, where the data may be financial, such as a sequence of stock prices; human-generated, such as a written text or a behavioural sequence; biological (DNA sequences); physical measurements and so on. In many of these applications very little, if anything, is known about the process that generates the data, and therefore it is hard to come up with reasonable assumptions. Moreover, achieving 0 asymptotic average error is often hopeless. For example, one can never hope to learn to predict accurately the probabilities of the stock market prices, even on long-term average; nor the probability distribution of the next letter of a human-written or other natural text, a problem that is directly linked to compressing such texts. This prompts a consideration of very general classes of environments \mathcal{C} , that would allow for some learning yet would also encompass as much as possible all the natural environments one tries to model.

One way to come up with such sets is considering changing environments. For instance, the data sequence may have a number of change points, such that between each two consecutive change points the sequence is generated by a relatively simple measure (e.g., i.i.d. or Markov), but the sequence of change points is essentially arbitrary, with the only constraint being a one on the frequency of changes. Another way may be to consider arbitrary additive trends: again, take a sequence generated by a measure from a relatively simple set and sum it up with another, which may be arbitrary except for a constraint

on how fast it changes.

These are just some of the ways of constructing large sets of measures \mathcal{C} . These methods do not come close to fully addressing the challenges arising in applications just mentioned. Here we do not concentrate on any particular example, but rather attempt to tackle the problem in its full generality.

The main result. Take any set \mathcal{C} of measures and an arbitrary predictor ρ . We show that there exists a Bayesian predictor, ν , such that its excess loss with respect to ρ is at most logarithmic:

$$L_n(\mu, \nu) \leq L_n(\mu, \rho) + 7 \log n + O(\log \log n)$$

for every $\mu \in \mathcal{C}$. Moreover, the prior of the Bayesian predictor can always be chosen to be discrete, that is,

$$\nu = \sum_{i=1}^{\infty} w_i \mu_i,$$

where $\mu_i \in \mathcal{C}$, and w_i are real weights. This in particular allows us to consider sets \mathcal{C} which may not be measurable. The constants in the $O(\cdot)$ term are small and are given explicitly; apart from absolute constants, there is also a linear dependence on the size of the alphabet $|\mathcal{X}|$.

This is a theoretical result. More steps remain to be made before real applications can be addressed. Perhaps the most important one is finding a general method for constructing the optimal prior whose existence is established in this work.

In addition, a lower bound is established, showing that there exists a set of measures \mathcal{C} and a measure ρ , such that for every Bayesian predictor ν whose prior is concentrated on \mathcal{C} , there exists a function $\theta(n) \rightarrow \infty$, there exist infinitely many time steps n_i and measures $\mu_i \in \mathcal{C}$ such that

$$L_{n_i}(\mu_i, \nu) \geq L_{n_i}(\mu_i, \rho) + \theta(n_i)$$

for all $i \in \mathbb{N}$.

Thus, there is an order-log n gap between the upper and the lower bounds.

Related work. Apart from the previously mentioned results on which this work builds, one should mention an alternative general approach to prediction, namely, prediction with expert advice Cesa-Bianchi und Lugosi (2006). Here it is assumed that the sequence one tries to predict is completely arbitrary, but, instead, one is given a set of predictors (or experts) \mathcal{C} to compete with. The relations between this setting and the one considered here have been analysed by Ryabko (2011). What is important to note is that for this problem in its full generality there is so far no generic method for constructing predictors to compete with an arbitrary set of experts \mathcal{C} . In particular, Ryabko (2016) shows that there are sets \mathcal{C} such that every Bayesian predictor has suboptimal asymptotic average regret. Note that such sets must necessarily be large (in particular, uncountable), while most of the work on expert advice concentrates on finite sets of experts \mathcal{C} or else on sets of experts satisfying some very specific properties. It remains open to find which properties of sets of experts are necessary and sufficient for any general algorithm (Bayesian or not) to be optimal.

2 Setup

Let \mathcal{X} be a finite set (the alphabet), and let

$$M := \log |\mathcal{X}|. \quad (1)$$

The notation $x_{1..n}$ is used for x_1, \dots, x_n . The symbol \mathbf{E}_μ denotes the expectation with respect to a measure μ . We consider (probability) measures on $(\mathcal{X}^\infty, \mathcal{F})$, where \mathcal{F} is the usual Borel sigma-field.

In general, a *Bayesian* predictor ν over a set \mathcal{C} is a measure $\int_{\mathcal{C}} \alpha dW(\alpha)$ where W is a measure over the set of all measures on $(\mathcal{X}^\infty, \mathcal{F})$, the latter being assumed endowed with the structure of a probability space Gray (1988). However, in this paper we shall only be dealing with Bayesian predictors with discrete priors, that is, with predictors of the form $\sum_{i=1}^{\infty} w_i \mu_i$, where $(w_i)_{i \in \mathbb{N}}$ are reals (that play the role of the distribution W above), and $\mu_i \in \mathcal{C}$, $i \in \mathbb{N}$. This allows us to avoid any measureability issues, in particular allowing \mathcal{C} to be non-measurable.

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

$$\begin{aligned} L_n(\mu, \rho) &:= \mathbf{E}_\mu \sum_{t=1}^{n-1} \sum_{a \in \mathcal{X}} \mu(x_t = a | x_{1..t-1}) \log \frac{\mu(x_t = a | x_{1..t-1})}{\rho(x_t = a | x_{1..t-1})} \\ &= \sum_{x_{1..n} \in \mathcal{X}^n} \mu(x_{1..n}) \log \frac{\mu(x_{1..n})}{\rho(x_{1..n})}. \end{aligned} \quad (2)$$

In words, we take the μ -expected (over data) cumulative (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. Here μ will be interpreted as the distribution generating the data.

3 Main result

The main result shows that the performance of any predictor can be matched by that of a Bayesian predictor with some prior, up to an additive $\log n$ term.

Theorem 1. *Let \mathcal{C} be any set probability measures on $(\mathcal{X}^\infty, \mathcal{F})$, and let ρ be another probability measure on this space, considered as a predictor. Then there is a discrete Bayesian predictor ν , that is, a predictor of the form $\sum_{k \in \mathbb{N}} w_k \mu_k$ where $\mu_k \in \mathcal{C}$ and $w_k \in [0, 1]$, such that for every $\mu \in \mathcal{C}$ we have*

$$L_n(\mu, \nu) - L_n(\mu, \rho) \leq 7 \log n + O(\log \log n), \quad (3)$$

where the constants in $O(\cdot)$ are small and are given in (24) using the notation defined in (1), (4), (18) and (25).

The proof (which follows below) uses the construction from Ryabko (2017) with a refined and added analysis that allows for rates extraction. The main ideas of the proof are as follows. First of all, a separate predictor is constructed to work on time steps 1 to n for each n ; these predictors are later summed up with weights to obtain the final predictor. Before going any further, note that constructing a predictor for each n must be done without forgetting the rest of the time indices n : in fact, taking a predictor that is minimax optimal for each n and summing these predictors up (with weights) for all $n \in \mathbb{N}$ may result in the worst possible predictor overall, and in particular, a one much worse than the predictor ρ given. An example of this behaviour is given in the proof of Theorem 2 (the lower bound). Thus, the measure ρ is used in an essential way when constructing a predictor for each of the time steps n . For each n , we consider a covering of the set \mathcal{X}^n with subsets, each of which is associated with a measure μ from \mathcal{C} . These latter measures are then those the prior is concentrated on (that is, they are summed up with weights). The covering is constructed as follows. The log-ratio function $\log \frac{\mu(x_{1..n})}{\rho(x_{1..n})}$, where ρ is the predictor whose performance we are trying to match, is approximated with a step function for each μ , and for each size of the step. The cells of the resulting partition are then ordered with respect to their ρ probability. The main part of the proof is then to show that not too many cells are needed to cover the set \mathcal{X}^n this way up to a small probability. Quantifying the “not too many” and “small” parts results in the final bound.

Proof. Define the weights

$$w_k := w/k \log^2(k+1) \quad (4)$$

where w is the normalizer such that $\sum_{k \in \mathbb{N}} w_k = 1$. Replacing ρ with $1/2(\rho + p)$ if necessary, where p is the i.i.d. measure with equal probabilities of outcomes, we shall assume (without loss of generality)

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

Thus, in particular,

$$L_n(\mu, \rho) \leq nM + 1 \text{ for all } \mu. \quad (6)$$

The first part of the proof is the covering construction from Ryabko (2017), which follows.

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

$$T_\mu^n := \left\{ x_{1..n} \in \mathcal{X}^n : \frac{\mu(x_{1..n})}{\rho(x_{1..n})} \geq \frac{1}{n} \right\}. \quad (7)$$

From Markov inequality, we obtain

$$\mu(\mathcal{X}^n \setminus T_\mu^n) \leq 1/n. \quad (8)$$

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. $U_k := \{u_k^i : i = 1..k\}$, where

$$u_k^i = \begin{cases} \left[-\frac{\log n}{n}, \frac{iM}{k} \right] & i = 1, \\ \left(\frac{(i-1)M}{k}, \frac{iM}{k} \right] & 1 < i < k, \\ \left(\frac{(i-1)M}{k}, M + \frac{1}{n} \right] & i = k. \end{cases} \quad (9)$$

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

$$T_{\mu, k, i}^n := \left\{ x_{1..n} \in \mathcal{X}^n : \frac{1}{n} \log \frac{\mu(x_{1..n})}{\rho(x_{1..n})} \in u_k^i \right\}. \quad (10)$$

Observe that, for every $\mu, k, n \in \mathbb{N}$, these sets constitute a partition of T_μ^n 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 (7) of T_μ^n , 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 $x_{1..n} \in T_{\mu, k, i}^n$ we have

$$\mu(x_{1..n}) \leq 2^{\frac{iM}{k}n+1} \rho(x_{1..n}). \quad (11)$$

For every $n, k \in \mathbb{N}$ and $i \in \{1..k\}$ consider the following construction. Define $m_1 := \max_{\mu \in \mathcal{C}} \rho(T_{\mu, k, i}^n)$ (since \mathcal{X}^n are finite all suprema are reached). Find any μ_1 such that $\rho(T_{\mu_1, k, i}^n) = m_1$ and let $T_1 := T_{\mu_1, k, i}^n$. For $l > 1$, let $m_l := \max_{\mu \in \mathcal{C}} \rho(T_{\mu, k, i}^n \setminus T_{l-1})$. If $m_l > 0$, let μ_l be any $\mu \in \mathcal{C}$ such that $\rho(T_{\mu_l, k, i}^n \setminus T_{l-1}) = m_l$, and let $T_l := T_{l-1} \cup T_{\mu_l, k, i}^n$; 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 $x_{1..n} \in T_{\mu_{l'}, k, i}^n$ and thus from (10) we get

$$2^{\frac{(i-1)M}{k}n - \log n} \rho(x_{1..n}) \leq \mu_{l'}(x_{1..n}). \quad (12)$$

Finally, define

$$\nu_{n, k, i} := \sum_{l=1}^{\infty} w_l \mu_l. \quad (13)$$

(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.) Finally, define the predictor ν as

$$\nu := \frac{1}{2} \sum_{n, k \in \mathbb{N}} w_n w_k \frac{1}{k} \sum_{i=1}^k \nu_{n, k, i} + \frac{1}{2} r, \quad (14)$$

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

$$\log \frac{\mu'(x_{1..n})}{\nu(x_{1..n})} \leq nM - \log w_n + 1 \quad \text{for all } x_{1..n} \in \mathcal{X}^n; \quad (15)$$

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 proceed to show that the ν is the predictor whose existence is claimed in the statement.

Introduce the notation

$$L_n|_A(\mu, \nu) := \sum_{x_{1..n} \in A} \mu(x_{1..n}) \log \frac{\mu(x_{1..n})}{\rho(x_{1..n})};$$

with this notation, for any set $A \subset \mathcal{X}^n$ we have

$$L_n(\mu, \nu) = L_n|_A(\mu, \nu) + L_n|_{\mathcal{X}^n \setminus A}(\mu, \nu).$$

First we want to show that for each $\mu \in \mathcal{C}$ for each fixed k, i the sets $T_{\mu, k, i}^n$ are covered by sufficiently few sets T_l , where “sufficiently few” is, in fact, exponentially many with the right exponent. By definition, for each n, i, k the sets $T_l \setminus T_{l-1}$ are disjoint (for different l) and have non-increasing (with l) ρ -probability. Therefore, $\rho(T_{l+1} \setminus T_l) \leq 1/l$ for all $l \in \mathbb{N}$. Hence, from the definition of T_l , we must also have $\rho(T_{\mu, k, i}^n \setminus T_l) \leq 1/l$ for all $l \in \mathbb{N}$. From the latter inequality and (11) we obtain $\mu(T_{\mu, k, i}^n \setminus T_l) \leq \frac{1}{l} 2^{\frac{iM}{k}n+1}$. Take $l_i := kn 2^{\frac{iM}{k}n+1}$ to obtain

$$\mu(T_{\mu, k, i}^n \setminus T_{l_i}) \leq \frac{1}{kn}. \quad (16)$$

Moreover, for every $i = 1..k$, for each $x_{1..n} \in T_{l_i}$, there is $l' \leq l_i$ such that the following chain holds

$$\begin{aligned} \nu(x_{1..n}) &\geq \frac{1}{2} w_n w_k \frac{1}{k} w_{kn 2^{\frac{iM}{k}n+1}} \mu_{l'}(x_{1..n}) \\ &\geq \frac{w^3}{4(M+1/n)^2 n^4 k^2 \log^3 n \log^2 k} 2^{-\frac{iM}{k}n} \mu_{l'}(x_{1..n}) \\ &\geq \frac{w^3}{4(M+1/n)^2 n^5 k^2 \log^3 n \log^2 k} 2^{-\frac{M}{k}n} \rho(x_{1..n}) \\ &= B_n 2^{-\frac{M}{k}n} \rho(x_{1..n}), \end{aligned} \quad (17)$$

where the first inequality is from (14) and (13) (with $l = l_i$), the second is by definition of w_l and $i \leq k$, in the third inequality we used (12), and the final equality introduces B_n defined as

$$B_n := \frac{w^3}{4(M+1/n)^2 n^5 k^2 \log^3 n \log^2 k}. \quad (18)$$

We have

$$L_n(\mu, \nu) = \sum_{i=1}^k L_n|_{T_{l_i}}(\mu, \nu) + L_n|_{\mathcal{X}^n \setminus \bigcup_{i=1}^k T_{l_i}}(\mu, \nu). \quad (19)$$

For the first term, from (17) we obtain

$$\begin{aligned} \sum_{i=1}^k L_n|_{T_{l_i}}(\mu, \nu) &\leq \sum_{i=1}^k L_n|_{T_{l_i}}(\mu, \rho) + Mn/k - \log B_n \\ &= L_n(\mu, \rho) - L_n|_{\mathcal{X}^n \setminus \cup_{i=1}^k T_{l_i}}(\mu, \rho) + Mn/k - \log B_n. \end{aligned} \quad (20)$$

For the second term in (19), we recall that $T_{\mu, k, i}^n$, $i = 1..k$ is a partition of T_μ^n , and decompose

$$\mathcal{X}^n \setminus \cup_{i=1}^k T_{l_i} \subseteq (\cup_{i=1}^k (T_{\mu, k, i}^n \setminus T_{l_i})) \cup (\mathcal{X}^n \setminus T_\mu^n). \quad (21)$$

Next, using (15) and an upper-bound for the μ -probability of each of the two sets in (21), namely, (16) and (8), we obtain

$$L_n|_{\mathcal{X}^n \setminus \cup_{i=1}^k T_{l_i}}(\mu, \nu) \leq (nM - \log w_n + 1) \frac{2}{n}. \quad (22)$$

Returning to (20), from Jensen's inequality one can show (see, e.g., Ryabko, 2010, equation 11) that, for any set $A \subset \mathcal{X}^n$,

$$-L_n|_A(\mu, \rho) \leq \mu(A) \log \rho(A) + 1/2.$$

Therefore, using (6), similarly to (22) we obtain

$$-L_n|_{\mathcal{X}^n \setminus \cup_{i=1}^k T_{l_i}}(\mu, \rho) \leq (nM + 1) \frac{2}{n} + \frac{1}{2}. \quad (23)$$

Combining (19) with (20), (22) and (23) we derive

$$L_n(\mu, \nu) \leq L_n(\mu, \rho) + Mn/k - \log B_n + 4M - \frac{2}{n}(\log w_n - 1) + 1/2; \quad (24)$$

setting

$$k := n / \log \log n \quad (25)$$

we obtain the statement of the theorem.

It remains to come back to (15) 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 (15). \square

4 Lower bound

In this section we establish a lower bound on being a Bayesian with the best prior. The bound leaves a significant gap with respect to the upper bound, but it shows that the regret of using the Bayesian predictor with the *best* prior for the given problem cannot be upper-bounded by a constant.

Theorem 2. *There exists a set of measures \mathcal{C} and a measure ρ , such that for every Bayesian predictor ν whose prior is concentrated on \mathcal{C} , there exists a function $\theta(n)$ which is non-decreasing and goes to infinity with n , there exist infinitely many time steps n_i and measures $\mu_i \in \mathcal{C}$ such that $L_{n_i}(\mu_i, \nu) - L_{n_i}(\mu_i, \rho) \geq \theta(n_i)$ for all $i \in \mathbb{N}$.*

Thus, the lower bound goes to infinity with n but may do so arbitrarily slow. This leaves a gap with respect to the $O(\log n)$ upper bound of Theorem 1.

Note also that this formulation is good enough to be the opposite of Theorem 1, because the formulation of the latter is strong: Theorem 1 says that *for every μ and for every n* (the regret is upper bounded), so, in order to counter that, it is enough to say that *there exists n and there exists μ* (such that the regret is lower bounded); Theorem 2 is, in fact, a bit stronger, since it establishes that there are infinitely many such n . However, it does not preclude that for every μ in \mathcal{C} the loss of the Bayesian is upper-bounded by a constant independent of n , while the loss of ρ is linear in n . This is, in fact, the case in the proof.

Proof. Let $\mathcal{X} := \{0, 1\}$. Let \mathcal{C} be the set of Dirac delta measures, that is, the measures each of which is concentrated on a single deterministic sequence, where the sequences are all sequences that are 0 from some n on. In particular, introduce $S_n := \{x_{1,2,\dots} \in \mathcal{X}^\infty : x_i = 0 \text{ for all } i > n\}$, $S := \cup_{n \in \mathbb{N}} S_n$. Let \mathcal{C}_n be the set of all measures μ such that $\mu(x) = 1$ for some $x \in S_n$ and let $\mathcal{C} := \cup_{n \in \mathbb{N}} \mathcal{C}_n$.

Observe that the set \mathcal{C} is countable. It is therefore, very easy to construct a (Bayesian) predictor for this set: enumerate it in any way, say $(\mu_k)_{k \in \mathbb{N}}$ spans all of \mathcal{C} , fix a sequence of positive weights w_k that sum to 1, and let

$$\nu := \sum_{k \in \mathbb{N}} w_k \mu_k. \quad (26)$$

Then $L_n(\mu_k, \nu) \leq -\log w_k$ for all $k \in \mathbb{N}$. That is, for every $\mu \in \mathcal{C}$ the loss of ν is upper-bounded by a constant: it depends on μ but not on the time index n . So, it is good for every μ for large n , but may be bad for some μ for (relatively) small n , which is what we shall exploit.

Observe that, since \mathcal{C} is countable, every Bayesian ν with its prior over \mathcal{C} must have, by definition, the form (26) for some weights $w_k \in [0, 1]$ and some measures $\mu_k \in \mathcal{C}$. Thus, we fix any Bayesian ν in this form.

Define ρ to be the Bernoulli i.i.d. measure with the parameter $1/2$. Note that

$$L_n(\mu, \rho) = n \quad (27)$$

for every n . This is quite a useless predictor; its asymptotic average error is the worst possible, 1. However, it is minimax optimal for every single time step n :

$$\inf_{\rho'} \sup_{\mu \in \mathcal{C}} L_n(\mu, \rho') = n,$$

where the inf is over all possible measures. This is why ρ is hard to compete with—and, incidentally, why being minimax optimal for each n separately may be useless.

For each $s \in \mathbb{N}$, let W_s be the weight that ν spends on the measures in the sets \mathcal{C}_k with $k < s$, and let M_s be the set of these measures:

$$W_s := \sum \{w_i : \exists k < s \text{ such that } \mu_i \in \mathcal{C}_k\},$$

and

$$M_s := \{\mu_i : \exists k < s \text{ such that } \mu_i \in \mathcal{C}_k\}.$$

By construction,

$$\lim_{s \rightarrow \infty} W_s = 1. \quad (28)$$

Next, for each $n \in \mathbb{N}$, let $U_n := S_{n+1} \setminus S_n$ (these are all the sequences in S_{n+1} with 1 on the n th position). Note that $\mu(U_n) = 0$ for each $\mu \in M_n$, while $|U_n| = 2^n$. From the latter equality, there exists $x_{1..n} \in \mathcal{X}^{n+1}$ and $\mu \in U_n \subset \mathcal{C}_{n+1}$ such that

$$\mu(x_{1..n} = 1) \text{ but } \nu(x_{1..n}) \leq 2^{-n}(1 - W_s).$$

This, (28) and (27) imply the statement of the theorem. \square

5 Conclusion and future work

The main result, Theorem 1, is the first one to show finite-time optimality of the Bayesian method for the prediction problem in full generality; or, perhaps, at this generality, for any learning problem. A number of important questions remain, both directly extending the result of this work and more general.

Lower bounds, necessity of the $\log n$ term. The first question is how sharp is the result. So far, the lower bound only shows that, for every prior, the Bayesian may suffer more than constant regret. The question whether the $\log n$ term is necessary remains open. If it is, one can ask what is the best constant in front. In the proof of the main result, the constant comes, first of all, from all the weights used in constructing the predictor, that is, from the prior. Each of the sums in (14) contributes one or two $\log n$. The outermost is perhaps (partially) removable with some version of the doubling trick: that is, instead of summing over all time steps $n \in \mathbb{N}$ one would only sum over some time steps, and reuse the predictors at remaining time steps. Yet, as the proof of Theorem 2 shows, some regret from summing up over different time steps is unavoidable. The rest are less clear how to optimize. Finally, one additional $\log n$ comes from the definition of the sets T_μ^n in (8), via the top line of (9). This one would be harder to remove, because the $\frac{1}{n}$ term is necessary in (22).

One could also ask the question of how important it is to optimize this constant. First of all, of course, it is only important if the $\log n$ term is at all necessary. But if it is necessary, then the constant is important, because the optimal loss is of order $\log n$ in some commonly studied special cases of \mathcal{C} , such as i.i.d. or Markov measures. (It is worth mentioning that the known optimal predictors in these cases Krichevsky (1993) are, in fact, Bayesian.)

Moreover, it may be worth trying to improve the bounds specifically for the case $L_n(\mu, \rho) = O(\log n)$, since in the opposite case it is not important.

Further generalizations. Some further natural and interesting generalizations are to different (or general) loss functions, as well as to infinite (countable or continuous) alphabets \mathcal{X} .

However, the most important direction for further research appears to be finding a general method of constructing a prior that results in an optimal predictor for an arbitrary class of measures \mathcal{C} . Another interesting question, mentioned in the introduction, is finding out under what conditions the Bayesian procedure, or indeed any other general method, is optimal for the non-realizable version of the problem; as discussed, some conditions are necessary, as shown in Ryabko (2016).

Finally, it is also interesting to find out to what extent the obtained result can be generalized to interactive learning problems, such as bandits, or, more generally, reinforcement learning.

References

- [Cesa-Bianchi und Lugosi 2006] CESA-BIANCHI, N. ; LUGOSI, G.: *Prediction, Learning, and Games*. Cambridge University Press, 2006. – ISBN 0521841089
- [Diaconis und Freedman 1986] DIACONIS, P. ; FREEDMAN, D.: On the Consistency of Bayes Estimates. In: *Annals of Statistics* 14 (1986), Nr. 1, S. 1–26
- [Freedman 1963] FREEDMAN, David A.: On the asymptotic behavior of Bayes estimates in the discrete case I. In: *The Annals of Mathematical Statistics* (1963), S. 1386–1403
- [Freedman 1965] FREEDMAN, David A.: On the asymptotic behavior of Bayes estimates in the discrete case II. In: *The Annals of Mathematical Statistics* 36 (1965), Nr. 2, S. 454–456
- [Gray 1988] GRAY, Robert M.: *Probability, Random Processes, and Ergodic Properties*. Springer Verlag, 1988
- [Krichevsky 1993] KRICHEVSKY, R.: *Universal Compression and Retrieval*. Kluwer Academic Publishers, 1993
- [Ryabko 2017] RYABKO, D.: Universality of Bayesian mixture predictors. In: *Proceedings of the 28th International Conference on Algorithmic Learning Theory (ALT'17)* Bd. 76. Kyoto, Japan : JMLR, 2017, S. 57–71

- [Ryabko 2010] RYABKO, Daniil: On Finding Predictors for Arbitrary Families of Processes. In: *Journal of Machine Learning Research* 11 (2010), S. 581–602
- [Ryabko 2011] RYABKO, Daniil: On the relation between realizable and non-realizable cases of the sequence prediction problem. In: *Journal of Machine Learning Research* 12 (2011), S. 2161–2180
- [Ryabko 2016] RYABKO, Daniil: Things Bayes can't do. In: *Proceedings of the 27th International Conference on Algorithmic Learning Theory (ALT'16)* Bd. 9925. Bari, Italy : Springer, 2016, S. 253–260