Journal of Machine Learning Research 20 (2019) 1-24 Submitted 7/18; Revised 4/19; Published 9/19

On Asymptotic and Finite-Time Optimality of Bayesian Predictors

Daniil Ryabko DANIIL@RYABKO.NET

Editor: Csaba Szepesvari

Abstract

The problem is that of sequential probability forecasting for finite-valued time series. The data is generated by an unknown probability distribution over the space of all one-way infinite sequences. Two settings are considered: the realizable and the non-realizable one. Assume first that the probability measure generating the sequence belongs to a given set $\mathcal{C}$ (realizable case), but the latter is completely arbitrary (uncountably infinite, without any structure given). It is shown that the minimax asymptotic average loss—which may be positive—is always attainable, and it is attained by a Bayesian predictor whose prior is discrete and concentrated on $\mathcal{C}$. Moreover, the finite-time loss of the Bayesian predictor is also optimal up to an additive $\log n$ term (where n is the time step). This upper bound is complemented by a lower bound that goes to infinity but may do so arbitrarily slow.

Passing to the non-realizable setting, let the probability measure generating the data be arbitrary, and consider the given set $\mathcal{C}$ as a set of experts to compete with. The goal is to minimize the regret with respect to the experts. It is shown that in this setting it is possible that all Bayesian strategies are strictly suboptimal even asymptotically. In other words, a sublinear regret may be attainable but the regret of every Bayesian predictor is linear.

A very general recommendation for choosing a model can be made based on these results: 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.

Keywords: sequence prediction, Bayesian prediction, complete-class theorems, minimax theorems

1. Introduction

[1] 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. The problem is considered in full generality; in particular, outcomes may exhibit arbitrary dependence. 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.

1. Some of the results in this paper were reported at ALT'17 and ALT'16 (Ryabko, 2017, 2016). All the non-asymptotic results (upper and lower bounds) in this paper are new.

The probabilities forecast by a predictor ρ having seen a sequence $x_1, \dots, x_n$ may be interpreted as conditional probabilities of x_{n+1} given $x_1, \dots, x_n$. Since the latter sequence is arbitrary, the predictor ρ defines a probability measure over the space of all one-way infinite sequences $\mathcal{X}^\infty$ (with the usual sigma-algebra generated by cylinder sets). Similarly, the mechanism that generates the data, that is, the observed sequence $x_1, \dots, x_n, \dots$, can be assumed probabilistic. Thus, predictors and the data-generating mechanism are objects of the same kind: probability measures of the space of one-way infinite sequences.

It is easy to see that without any model for the data nor for the predictor it is impossible to do meaningful inference for this task. A model can be seen as a set $\mathcal{C}$ of distributions over the space of one-way infinite sequences. There are two common approaches to use a model. The first one is to assume that the probability measure generating the sequence belongs to the model set $\mathcal{C}$. The goal is then to have a predictor whose (time-average) error, or *loss*, decreases (to zero), as fast as possible, as more and more data is revealed. The other approach is to suppose that the data may be an arbitrary individual sequence, and the model set $\mathcal{C}$ is considered a set of experts. The goal then is to have a predictor whose *regret* with respect to using the best expert in the set for the given sequence is as small as possible, and at least sublinear (average regret goes to zero as more data is revealed).

Here I would like to argue in favour of a third approach. It consists in making the model set $\mathcal{C}$ large enough to include the distribution that generates the data, even if it means that the best achievable asymptotic loss is strictly positive. In particular, whichever aspect of the data-generation process is completely unknown can be modelled by an arbitrary sequence.

Throughout the paper, the loss is chosen to be the Kullback-Leibler divergence (expected log-loss) mainly for the sake of convenience, but also reflecting its importance in applications. To fix ideas, consider a simple and familiar example (perhaps first considered by Willems 1996 in a somewhat different formulation) in which the best asymptotic average loss is non-zero: a piece-wise i.i.d. sequence. The sequence consists of segments, in each of which the data is i.i.d., but the distribution in each of the segments is otherwise arbitrary, and the sequence of points at which one segment ends and the next starts is completely arbitrary as well. Thus, one can learn the distribution within each segment but cannot predict when or to what the next change will be. If the frequency of these change points does not vanish then the minimax asymptotic average error is positive. This example is considered in more detail in Section 4.

Two results are presented to support the suggested approach to the general problem of sequential prediction.

The **first result** shows that, no matter how big a model class $\mathcal{C}$ is, if the distribution generating the data belongs to $\mathcal{C}$ then the optimal minimax asymptotic average performance, even if positive, is always achievable, and it can be attained by a Bayesian combination of countably many distributions in $\mathcal{C}$. One cannot, in general, say anything about the speed of convergence of the average loss, as it depends on $\mathcal{C}$ and may be arbitrarily slow for some classes $\mathcal{C}$. However, for any predictor ρ whatsoever, its performance can be matched by a Bayesian predictor up to an additive $O(\log n)$ term, with only rather small constants hidden inside the $O()$ term.

This means that, if the data-generating mechanism belongs to the model class, then no matter how big the latter is, the statistician knows where to start: it suffices to find the right prior distribution, and then the inference can be made by evaluating the posterior

with respect to the observed data. The performance guarantees would be pointwise (for every distribution in $\mathcal{C}$) rather than Bayesian (with prior probability 1 or in expectation), and it is possible to achieve nearly the best speed of decrease of the average loss. This adds a perspective to the classical results (Freedman, 1963, 1965; Diaconis and Freedman, 1986) that show that there are priors with which Bayesian inference is inconsistent even for problems on i.i.d. data (although, in that latter case, for infinite alphabets).

In some more detail, consider an arbitrary set $\mathcal{C}$ of probability distributions over one-way infinite sequences. Let $L_n(\mu, \nu)$ be the cumulative expected log loss (KL divergence) of ν used to predict μ, up to time n. A Bayesian predictor is a predictor of the form $\int \alpha dW(\alpha)$ where W is a probability distribution over $\mathcal{C}$ (a prior), that is, $W(\mathcal{C}) = 1$. The prediction is simply by evaluating the posterior over the given sequence $x_1, \dots, x_n$. If the prior is discrete (and so the integral is a sum) then we call such a predictor discrete Bayesian. While the general definition requires the structure of a probability space on $\mathcal{C}$ (or at least some subset of $\mathcal{C}$), for discrete Bayesian predictors no such structure is required. For the main result below, discrete Bayesian predictors are sufficient, so we do not need to worry about the measurability of $\mathcal{C}$. The main result is the following.

Theorem 1 *For every set $\mathcal{C}$ of probability measures and for every predictor ρ there is a discrete Bayesian predictor ν such that for every $\mu \in \mathcal{C}$ we have*

$$L_n(\mu, \nu) \leq L_n(\mu, \rho) + 8 \log n + O(\log \log n).$$

There are no assumptions whatsoever needed for this result to hold. As mentioned above, the set $\mathcal{C}$ is not even required to be measurable, and the distributions in $\mathcal{C}$ may be completely arbitrary (no restrictions on dependence, memory, etc.). The constants in $O()$ are small and, besides absolute constants, only include linear dependence on the alphabet size $|\mathcal{X}|$.

To the author's knowledge, the upper bound presents the first non-asymptotic result in a setting of this generality.

A lower bound is also obtained, which takes the form of $\theta(n)$ where θ increases to infinity, but may do so arbitrarily slow. Thus, there exist a set $\mathcal{C}$ for which by choosing to be Bayesian a statistician would suffer a more-than-constant cumulative regret. This is made precise in Section 3.2.

It is also useful to put this result into the context of the **decision theory**; from this point of view, it can be seen (albeit with some important differences) as a *complete-class theorem* for asymptotic average performance: it shows that the class of Bayes strategies is always essentially complete. Decision theory typically deals with single-shot games and general losses (here we only consider expected average KL divergence), and Wald's complete class theorem and its generalizations require a number of conditions to hold. Other important differences, besides the absence of any conditions in Theorem 1, include the fact that in our case all strategies are inadmissible and one cannot speak about minimal complete classes. To further understand the game-theoretic side of the main result, it is useful to consider the minimax asymptotic average loss $V_{\mathcal{C}}$, defined as

$$\inf_{\rho} \sup_{\mu \in \mathcal{C}} \limsup_{n \to \infty} \frac{1}{n} L_n(\mu, \rho).$$

It can be shown that the theorem above implies that this value can be attained by a Bayesian predictor with a discrete prior. Besides providing a complete-class theorem for

(asymptotic) prediction, this result can also be seen as a partial minimax theorem: for every set of strategies of the opponent, the statistician has a minimax strategy. For the "classical" case of $V_{\mathcal{C}} = 0$, that is, for the case when the average error of the best predictor vanishes, an analogous asymptotic result was obtained by Ryabko (2010). We shall see that the general case presents principled differences (not limited to the proofs). In particular, as mentioned above, in the case $V_{\mathcal{C}} > 0$ all strategies are inadmissible and there are no minimal complete classes, while if $V_{\mathcal{C}} = 0$ then every Bayes rule that achieves it is admissible and constitutes a minimal complete class. These results and implications are considered in Section 5.

The **second result** presented concerns regret minimization. That is, we turn the tables, and assume that the data-generating mechanism is unrestricted, in particular it can be an arbitrary (deterministic) sequence. The set of probability measures $\mathcal{C}$ is now the set of experts, and the goal of the statistician is to find a strategy that is as good as the best one in $\mathcal{C}$, for the given data. Now that the distribution generating the data may not be in $\mathcal{C}$, it may happen that every combination of the experts has strictly positive asymptotic average regret, even though 0 asymptotic average regret is achievable. The formal result concerns Bayesian combinations, but it is argued in Section 6 that it applies more generally.

Theorem 2 *There exists a measurable set $\mathcal{C}$ such that every Bayesian predictor with prior over $\mathcal{C}$ has a linear regret with respect to $\mathcal{C}$, while there exists a predictor ρ whose regret is sublinear.*

Putting the **two results together**, in the realizable case of the sequence prediction problem the best asymptotic performance can always be attained by a Bayesian strategy, while in the non-realizable case it is possible that all Bayesian strategies are strictly sub-optimal. We can therefore make 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.

Thus, the results presented invite a reconsideration of the familiar trade-off in model selection: the model must be large enough to describe the data but small enough to be learnable. In reality, the first part is often eschewed in favour of the second: the model does not include the data-generating mechanism, but allows for learning. In the context of sequence prediction, learnability of the model is usually understood as allowing for the average error or regret of the predictor to go to zero. In light of the presented results, it can be suggested to change the preference to the other side of the trade-off, namely, make the model large enough to include the data, since having a linearly increasing loss does not preclude one from finding the best predictor. On the other hand, not having the data-generating mechanism inside the model may mean that there is no optimal Bayesian predictor, or, as argued below in Section 6, any combination of predictors in the model class. Of course, this recommendation is not relevant for the specific cases where one knows how to find an optimal predictor that does not have such a form.

1.1. Related Work

The literature on sequential prediction is vast and spread between several fields: nonparametric statistics, machine learning, information theory and econometrics. While an ex-

haustive survey is beyond the scope of this work, some related results and some important differences in the settings are worth mentioning. Most of the literature deals with specific families of processes $\mathcal{C}$, and almost exclusively with the case $V_{\mathcal{C}} = 0$. For the latter case, an asymptotic analogue of Theorem 1 was established by Ryabko (2010). An important exception from the $V_{\mathcal{C}} = 0$ case is the problem of prediction of processes with abrupt changes mentioned above, even though the formulation considered previously does not introduce the value $V_{\mathcal{C}}$, since the loss of the predictor is measured with respect to the best predictor that knows the actual changes (expert-advise formulation); see, e.g., (Gyorgy et al., 2012) and references. Perhaps the most general specific classes of processes considered in the literature are those of stationary ergodic distributions (Ryabko, 1988; Morvai et al., 1997; Gyorfi and Ottucsak, 2007).

The problem of regret minimization in prediction is usually considered in a slightly different setting, called prediction with expert advice; an overview is provided by Cesa-Bianchi and Lugosi (2006). One of the key differences is that the loss is measured with respect to actual outcomes rather than probabilities, making the setting non-probabilistic even though the predictions are usually randomized. An attempt to put the two settings under the same formulation and thus making the results directly comparable has been taken by Ryabko (2011), which also poses the question that Theorem 2 answers.

The econometrics literature on the subject concerns the $V_{\mathcal{C}} = 0$ case (with a different loss) and studies sufficient conditions on the existence of consistent predictors (loss goes to 0 asymptotically), mostly in the Bayesian setting, which means that the results are with prior probability 1 and not pointwise; see (Kalai and Lehrer, 1994; Noguchi, 2015) and references.

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, 1976-1979; Haussler, 1997).

1.2. Organization

The rest of the paper is organized as follows. The next section introduces notation and definitions. Section 3 presents the main asymptotic and finite-time results on the existence of Bayesian prediction for minimizing loss, with Section 3.2 providing the lower bound. Section 4 provides examples of classes for which $V_{\mathcal{C}} > 0$, illustrating the main findings. Decision-theoretic interpretations of the asymptotic result are given in section 5. Section 6 presents the impossibility result for regret minimization, and Section 7 concludes.

2. Notation and Definitions

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

$$M := \log |\mathcal{X}|. \tag{1}$$

The notation $x_{1..n}$ is used for $x_1, \dots, x_n$. $\mathbb{N}$ is the set of naturals without 0; for a finite set A denote $|A|$ its cardinality, and $A^* := \cup_{k\in\mathbb{N}} A^k$. All logarithms are base 2. We use $\mathbf{E}_\mu$ for expectation with respect to a probability measure μ. We consider probability measures on $(\mathcal{X}^\infty, \mathcal{F})$, where $\mathcal{F}$ is the usual Borel sigma-field generated by cylinder sets (see below). The set of probability measures over $(\mathcal{X}^\infty, \mathcal{F})$ is denoted $\mathcal{P}(\mathcal{X}^\infty, \mathcal{F})$ or $\mathcal{P}$ for short.

A probability measure ρ is a *discrete Bayesian predictor* with a prior over $\mathcal{C}$ if $\rho = \sum_{k\in\mathbb{N}} w_k\mu_k$, for some measures $\mu_k \in \mathcal{C}$, $k \in \mathbb{N}$, where w_k are real weights that sum to 1. The latter weights can be seen as a prior distribution over (a subset of) $\mathcal{C}$.

In Section 6 we need a more general definition of Bayesian predictors, which allows for non-discrete priors, and thus require a structure of the probability space on $\mathcal{P}$. Here we shall define it in a standard way, following Gray (1988). Recall that the sigma-algebra $\mathcal{F}$ of the space of infinite sequences $(X^\infty, \mathcal{F})$ can be generated by the countable set $(B_i)_{i\in\mathbb{N}}$ of cylinders, $B_i := \{\mathbf{x} \in \mathcal{X}^\infty : x_{1..|b_i|} = b_i\}$ where the words b_i take all possible values in $|\mathcal{X}|^*$. Next, consider the countable set of sets $\{\nu \in \mathcal{P} : \nu(B_i) \in u\}$, where $u \subset [0, 1]$, obtained by taking all $B_i, i \in \mathbb{N}$ and all intervals u with rational endpoints. This set generates a sigma-algebra $\mathcal{F}'$. Denoting $\mathcal{P}'$ the set of all probability measures over $\mathcal{P}$ we obtain the measurable space $(\mathcal{P}', \mathcal{F}')$.

Associated with any probability measure $W \in \mathcal{P}'$ there is a probability measure $\rho_W \in \mathcal{P}$ defined by $\rho_W = \int_P \alpha dW(\alpha)$ (*barycentre*, in the terminology of Gray (1988); see the latter work for a detailed exposition). A measure $\rho \in \mathcal{P}$ is *Bayesian* for a set $\mathcal{C} \subset \mathcal{P}$ if $\rho = \rho_W$ for some $W \in \mathcal{P}'$ such that there exists a measurable set $\bar{\mathcal{C}} \subset \mathcal{C}$ with $W(\bar{\mathcal{C}}) = 1$. (The subset $\bar{\mathcal{C}}$ is necessary since $\mathcal{C}$ may not be measurable.) The reason behind the name is that W can be seen as a prior over $\mathcal{C}$, and ρ_W as a predictor whose predictions are simply by evaluating the posterior distribution $\rho_W(\cdot|x_1, \dots, x_n)$ where $x_1, \dots, x_n$ are observations up to time n.

2.1. Loss

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

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

In words, we take the μ-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. Here μ will be interpreted as the distribution generating the data.

The asymptotic average expected loss is then defined as

$$\bar{L}(\nu, \rho) := \limsup_{n\to\infty} \frac{1}{n} L_n(\nu, \rho),$$

where the upper limit is chosen so as to reflect the worst performance over time. One can define the worst-case performance of a strategy ρ by

$$\bar{L}(\mathcal{C}, \rho) := \sup_{\mu\in\mathcal{C}} \bar{L}(\mu, \rho)$$

and the minimax value by

$$V_{\mathcal{C}} := \inf_{\rho\in\mathcal{P}} \bar{L}(\mathcal{C}, \rho). \tag{3}$$

Some examples of calculating the latter value for different sets $\mathcal{C}$ are considered in Section 4; the most common case in the literature is $V_{\mathcal{C}} = 0$, which is the case, for example, if $\mathcal{C}$ is the set of all i.i.d. or all stationary distributions.

2.2. Regret

Switching the roles, assume that the set of strategies of the opponent is unrestricted; the set of probability measures $\mathcal{C} \subset \mathcal{P}(X^\infty, \mathcal{F})$ is now the set of experts, and the goal of the statistician is to find a strategy that is as good as the best one in $\mathcal{C}$, for the given data. Thus, we are interested in the (asymptotic) *regret*

$$\bar{R}^\nu(\mu, \rho) := \limsup_{n\to\infty} \frac{1}{n} \left[L_n(\nu, \rho) - L_n(\nu, \mu)\right],$$

of using ρ as opposed to μ on the data generated by ν. The goal is to find ρ that minimizes the worst-case (over data) regret with respect to the best expert from the given set $\mathcal{C}$:

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

Note than in the expert-advice literature the regret is typically defined on finite sequences of length n, thus allowing both the experts and the algorithms to depend on n explicitly.

Similarly to $V_\mathcal{C}$, we can now define the value

$$U_\mathcal{C} := \inf_{\rho\in\mathcal{P}} R(\mathcal{C}, \rho),$$

which is the worst-case asymptotic average regret with respect to the set of experts $\mathcal{C}$.

In view of the (negative) result that is obtained for regret minimization, here we are mostly concerned with the case $U_\mathcal{C} = 0$.

3. Realizable Case: Minimizing Loss, Optimality of Bayes Rules

The main result (Theorem 1) shows that for any predictor ρ there is a Bayesian predictor that is as good as ρ, up to a $O(\log n)$ loss. It follows (Corollary 2) that the minimax loss is always achievable and is achieved by a discrete Bayesian predictor — without any assumptions on $\mathcal{C}$.

Theorem 1 (upper bound on the best Bayesian) *Let $\mathcal{C}$ be any set of 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*

$$L_n(\mu, \nu) - L_n(\mu, \rho) \le 8 \log n + O(\log\log n), \tag{4}$$

where the constants in $O(\cdot)$ are small and are given in (26) *using the notation defined in* (1), (6), (20) *and* (27). *The dependence on the alphabet size, M, is linear ($M \log\log n$) and the rest of the constants are universal.*

The proof is given after the corollary.

Corollary 2 (asymptotic optimality of Bayesian predictors) *For any set $\mathcal{C}$ of probability measures on $(\mathcal{X}^\infty, \mathcal{F})$, there exist a discrete Bayesian predictor φ such that*

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

Proof Note that the statement does not immediately follow from (4), because ρ in (4) may be such that $\sup_{\mu\in\mathcal{C}} \bar{L}(\mu,\rho) > V_{\mathcal{C}}$. Thus, let $\gamma_j > V_{\mathcal{C}}$, $j \in \mathbb{N}$ be a non-increasing sequence such that $\lim_{j\to\infty}\gamma_j = V_{\mathcal{C}}$. By the definition (3) of $V_{\mathcal{C}}$, it is possible to find a sequence $\rho_j \in \mathcal{P}$ such that $\bar{L}(\mathcal{C},\rho_j) \le \gamma_j$ for all $j \in \mathbb{N}$. From Theorem 1 we conclude that for each ρ_j, $j \in \mathbb{N}$ there is a probability measure ν_j of the form $\sum_{k\in\mathbb{N}} w'_k\mu_k$, where $\mu_k \in \mathcal{C}$ such that $\bar{L}(\mathcal{C},\nu_j) \le \bar{L}(\mathcal{C},\rho_j)$. It remains to define $\varphi := \sum_{j\in\mathbb{N}} w_j\nu_j$, where w_j are positive and sum to 1. Clearly, φ is a discrete Bayesian predictor. Let us show that for every $j \in \mathbb{N}$ it satisfies

$$\bar{L}(\mathcal{C},\varphi) \le \bar{L}(\mathcal{C},\rho_j). \tag{5}$$

Indeed, for every $\mu \in \mathcal{C}$ and every $j \in \mathbb{N}$

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

so that $\bar{L}(\mu,\varphi) \le \bar{L}(\mu,\nu_j) \le \bar{L}(\mu,\rho_j) \le \gamma_j$, establishing (5). Finally, recall that $\gamma_j \to V_{\mathcal{C}}$ to obtain the statement of the corollary. ■

3.1. Proof of Theorem 1

Before giving the proof of the theorem, let us briefly expose the main ideas behind it. Assume for a moment that, for each $\mu \in \mathcal{C}$, the limit $\lim_{n\to\infty}\frac{1}{n}\log\frac{\mu(x_{1..n})}{\rho(x_{1..n})}$ exists for μ-almost all $\mathbf{x} = x_1,\dots,x_n,\dots \in \mathcal{X}^\infty$, where ρ is the predictor given to compare to. Then we could define (μ-almost everywhere) the function $f_\mu(\mathbf{x})$ whose value at $\mathbf{x}$ equals this limit. Let us call it the "log-density" function. What we would be looking for thence is to find a countable dense subset of the set of log-densities of all probability 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. In reality, what we do instead in the proof 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 $\mathcal{C}$. Before going further, note that constructing a predictor for each n does not mean constructing the best predictors up to this time step: in fact, taking a predictor that is minimax optimal up to n, 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 3 (the lower bound). The objective for each n is different, and it is to approximate the measure ρ up to this time step with measures from $\mathcal{C}$. 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.

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 only obtains the asymptotic loss of $2V_{\mathcal{C}}$ for the mixture predictor presented there. It is not a problem for the case $V_{\mathcal{C}} = 0$, but of course is not what we want in the general case. The reason behind this problem is that for the construction in the proof of Ryabko (2010) 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 [of Theorem 1.] Define the weights w_k as follows: $w_1 := 1/2$, and, for $k > 1$

$$w_k := w/k \log^2 k, \tag{6}$$

where w is the normalizer such that $\sum_{k\in\mathbb{N}} w_k = 1$. Replacing ρ with $1/2(\rho+\delta)$ if necessary, where δ is the i.i.d. probability measure with equal probabilities of outcomes, i.e. $\delta(x_{1..n}) = M^{-1/n}$ for all $n \in \mathbb{N}, x_{1..n} \in \mathcal{X}^n$, we shall assume, without loss of generality,

$$-\log \rho(x_{1..n}) \le nM + 1 \text{ for all } n \in \mathbb{N} \text{ and } x_{1..n} \in \mathcal{X}^n. \tag{7}$$

The replacement is without loss of generality as it adds at most 1 to the final bound (to be accounted for). Thus, in particular,

$$L_n(\mu, \rho) \le nM + 1 \text{ for all } \mu. \tag{8}$$

The first part of the proof is the following covering construction.

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

$$T^n_\mu := \left\{ x_{1..n} \in \mathcal{X}^n : \frac{\mu(x_{1..n})}{\rho(x_{1..n})} \ge \frac{1}{n} \right\}. \tag{9}$$

From Markov inequality, we obtain

$$\mu(\mathcal{X}^n \backslash T^n_\mu) \le 1/n. \tag{10}$$

For each $k > 1$ 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

$$u^i_k = \begin{cases} \left[-\frac{\log n}{n}, \frac{iM}{k}\right] & i = 1, \\ \left(\frac{(i-1)M}{k}, \frac{iM}{k}\right] & 1 < i < k, \\ \left(\frac{(i-1)M}{k}, M + \frac{1}{n}\right] & i = k. \end{cases} \tag{11}$$

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 $\mu \in \mathcal{C}$, $n, k > 1$, $i = 1..k$ define the sets

$$T^n_{\mu,k,i} := \left\{ x_{1..n} \in \mathcal{X}^n : \frac{1}{n} \log \frac{\mu(x_{1..n})}{\rho(x_{1..n})} \in u^i_k \right\}. \tag{12}$$

Observe that, for every $\mu \in \mathcal{C}$, $k, n > 1$, 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 (9) 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 (7). In particular, from this definition, for all $x_{1..n} \in T^n_{\mu,k,i}$ we have

$$\mu(x_{1..n}) \leq 2^{\frac{iM}{k} n + 1} \rho(x_{1..n}). \tag{13}$$

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 $\mathcal{X}^n$ 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 (12) we get

$$2^{\frac{(i-1)M}{k} n - \log n} \rho(x_{1..n}) \leq \mu_{l'}(x_{1..n}). \tag{14}$$

Finally, define

$$\nu_{n,k,i} := \sum_{l=1}^{\infty} w_l \mu_l. \tag{15}$$

(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

$$\nu := \frac{1}{2} \sum_{n,k \in \mathbb{N}} w_n w_k \frac{1}{k} \sum_{i=1}^{k} \nu_{n,k,i} + \frac{1}{2} r, \tag{16}$$

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

$$\log \frac{\mu'(x_{1..n})}{\nu(x_{1..n})} \leq nM - \log w_n + 1 \ \text{ for all } x_{1..n} \in \mathcal{X}^n; \tag{17}$$

this and the stronger statement (7) for ν can be obtained analogously to the latter inequality in the case the i.i.d. measure δ is in $\mathcal{C}$; otherwise (since we need to define ν as a combination of probability 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.

Next, let us show that the measure 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 $T^n_{\mu,k,i}$ 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, $\rho(T_{l+1}\setminus T_l) \le 1/l$ for all $l \in \mathbb{N}$. Hence, from the definition of T_l, we must also have $\rho(T^n_{\mu,k,i}\setminus T_l) \le 1/l$ for all $l \in \mathbb{N}$. From the latter inequality and (13) we obtain

$$\mu(T^n_{\mu,k,i}\setminus T_l) \le \frac{1}{l} 2^{\frac{iM}{k}n+1}.$$

Take $l_i := \lceil kn 2^{\frac{iM}{k}n+1} \rceil$ to obtain

$$\mu(T^n_{\mu,k,i}\setminus T_{l_i}) \le \frac{1}{kn}. \tag{18}$$

Moreover, for every $i = 1..k$, for each $x_{1..n} \in T_{l_i}$, there is $l' \le l_i$ such that $x_{1..n} \in T^n_{\mu_{l'},k,i}$ and thus the following chain holds

$$\begin{aligned}
\nu(x_{1..n}) &\ge \frac{1}{2} w_n w_k \frac{1}{k} \nu_{n,k,i} \ge \frac{1}{2} w_n w_k \frac{1}{k} w_{kn\,2^{\frac{iM}{k}n+1}} \mu_{l'}(x_{1..n}) \\
&\ge \frac{w^3}{4n^2k^3 \log^2 n \log^2 k (\log k + \log n + 1 + nMi/k)^2} 2^{-\frac{iM}{k}n} \mu_{l'}(x_{1..n}) \\
&\ge \frac{w^3}{4(M+1)^2 n^4 k^3 \log^2 n \log^2 k} 2^{-\frac{iM}{k}n} \mu_{l'}(x_{1..n}) \\
&\ge \frac{w^3}{4(M+1)^2 n^5 k^3 \log^2 n \log^2 k} 2^{-\frac{M}{k}n} \rho(x_{1..n}) = B_n 2^{-\frac{M}{k}n} \rho(x_{1..n}),
\end{aligned} \tag{19}$$

where the first inequality is from (16), the second from (15) with $l = l_i$, the third is by definition of w_l, the fourth uses $i \le k$ for the exponential term, as well as $(\log n + \log k) \le n - 1$ for $n \ge 3$, which will be justified by the choice of k in the following (27), the fifth inequality uses (14), and the final equality introduces B_n defined as

$$B_n := \frac{w^3}{4(M+1)^2 n^5 k^3 \log^2 n \log^2 k}. \tag{20}$$

We have

$$L_n(\mu,\nu) = \left(\sum_{i=1}^k L_n|_{T_{l_i}}(\mu,\nu)\right) + L_n|_{\mathcal{X}^n\setminus\cup_{i=1}^k T_{l_i}}(\mu,\nu). \tag{21}$$

For the first term, from (19) we obtain

$$\sum_{i=1}^k L_n|_{T_{l_i}}(\mu,\nu) \le \sum_{i=1}^k L_n|_{T_{l_i}}(\mu,\rho) + Mn/k - \log B_n$$
$$= L_n(\mu,\rho) - L_n|_{\mathcal{X}^n\setminus\cup_{i=1}^k T_{l_i}}(\mu,\rho) + Mn/k - \log B_n. \tag{22}$$

For the second term in (21), we recall that $T^n_{\mu,k,i}$, $i = 1..k$ is a partition of T^n_μ, and decompose

$$\mathcal{X}^n \setminus \cup_{i=1}^k T_{l_i} \subseteq \left(\cup_{i=1}^k (T^n_{\mu,k,i} \setminus T_{l_i})\right) \cup (\mathcal{X}^n \setminus T^n_\mu). \tag{23}$$

Next, using (17) and an upper-bound for the μ-probability of each of the two sets in (23), namely, (18) and (10), as well as $k \ge 1$, we obtain

$$L_n|_{\mathcal{X}^n\setminus\cup_{i=1}^k T_{l_i}}(\mu,\nu) \le (nM - \log w_n + 1)\frac{2}{n}. \tag{24}$$

Returning to (22), from Jensen's inequality one can show (see, e.g., Ryabko, 2010, equation 11) that, for any set $A \subset \mathcal{X}^n$,

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

Therefore, using (8), similarly to (24) we obtain

$$-L_n|_{\mathcal{X}^n\setminus\cup_{i=1}^k T_{l_i}}(\mu,\rho) \le (nM+1)\frac{2}{n} + \frac{1}{2}. \tag{25}$$

Combining (21) with (22), (24) and (25) we derive

$$L_n(\mu,\nu) \le L_n(\mu,\rho) + Mn/k - \log B_n + 4M - \frac{2}{n}(\log w_n - 1) + 1/2; \tag{26}$$

setting

$$k := \lceil n/\log\log n\rceil \tag{27}$$

we obtain the statement of the theorem.

It remains to come back to (17) 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

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

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

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

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

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

for every $n \in \mathbb{N}$ and every $x_{1..n} \in A_n$, establishing (17). ■

3.2. Lower Bound

In this section we establish a lower bound on being a Bayesian, complementing the upper bound of Theorem 1. The bound leaves a significant gap with respect to the upper bound, but it shows that the regret of using the Bayesian predictor even with the *best* prior for the given set $\mathcal{C}$ cannot be upper-bounded by a constant.

Theorem 3 *There exists a measurable set of probability measures $\mathcal{C}$ and a probability 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. Effectively, the theorem compares the regret of the (best) Bayesian with respect to using the best predictor for $\mathcal{C}$ — but not with using the best predictor for each $\mu \in \mathcal{C}$, which is always μ itself.

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 3 is, in fact, a bit stronger, since it establishes that there are infinitely many such n. However, it does not preclude that for every fixed measure μ in $\mathcal{C}$ the loss of the Bayesian is upper-bounded by a constant independent of n (but dependent on μ), while the loss of ρ is linear in n. This is actually the case in the proof.

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

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

$$\nu := \sum_{k\in\mathbb{N}} w_k \mu_k. \tag{28}$$

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 (28) 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

$$L_n(\mu, \rho) = n \tag{29}$$

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 probability 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,

$$\lim_{s\to\infty} W_s = 1. \tag{30}$$

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 nth position). Note that $\mu(U_n) = 0$ for each $\mu \in M_n$, while $|U_n| = 2^n$. From the latter equality, there exists $x_{1..n} \in \mathcal{X}^n$ and $\mu \in U_n \subset S_{n+1}$ such that

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

This, (30) and (29) imply the statement of the theorem. ■

4. Examples

The main object of interest here are sets $\mathcal{C}$ for which $V_{\mathcal{C}} > 0$, as well as corresponding Bayesian minimax predictors. Various examples of Bayesian predictors for sets $\mathcal{C}$ for which $V_{\mathcal{C}} = 0$ are analysed by Ryabko (2010), so we do not consider this case here.

4.1. Typical Bernoulli 1/3 Sequences

We start with an example which is somewhat artificial, but comes up as a component in more realistic cases. 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: $\bar{L}(\mathcal{C}, \delta_{1/3}) = h(1/3)$, 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 $\mathcal{C}$, which 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 $n \ge n_\varepsilon(\mathbf{x})$, the average number of 1s in $x_{1..n}$ is within ε of 1/3. Fix the sequence of indices $n_j := 2^j$, $j \in \mathbb{N}$ and the sequence of thresholds $\varepsilon_l := 2^{-l}$. For each j let $S'^l_j \subset S$ 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^l_j of ${S'}^l_j$ such that for each $\mathbf{x}' \in {S'}^l_j$ there is $\mathbf{x} \in S$ such that $x'_{1..n_j} = x_{1..n_j}$. This is possible, since the set $\mathcal{X}^{n_j}$ is finite. Now for each $\mathbf{x} \in S^l_j$ take the corresponding measure $\mu_{\mathbf{x}} \in \mathcal{C}$ and attach to it the weight $w_l w_j / |S^l_j|$, 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^l_j 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.

4.2. Processes with Abrupt Changes

Start with a family of distributions S, for which we have a good predictor: for example, take S to be the set B of all Bernoulli i.i.d. processes, or, more generally, a set for which $V_S = 0$. The family $\mathcal{C}_\alpha$ 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 $\limsup_{i\to\infty} \frac{1}{n}\{i : n_i < n\}$ 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 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 by Willems (1996) for the case $S = B$, with the goal of minimizing the regret with respect to 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 average loss up to time n is $\frac{1}{n}(\frac{1}{2}\log n + 1)$, for the sequence $1..n_t$ it is $\frac{1}{n_t}\sum_{i=1}^{t}(\frac{1}{2}\log(n_i - n_{i-1}) + 1)$, where $n_0 := 1$. By Jensen's inequality, this sum is maximized when all the segments $n_i - n_{i-1}$ are of the same length, $1/\alpha$, so the total average loss is upper-bounded by $\alpha(1 - \frac{1}{2}\log\alpha)$. This value is also attainable, and thus gives $V_{\mathcal{C}_{B,\alpha}}$. 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) 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 $V_S = 0$).

4.3. 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 a 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 $x_i \in y$. Here again 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: such as a price drop; or a rain. At other times, the process is unpredictable: it can, again, be an arbitrary 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 $V_{\mathcal{C}} \in (0, 1)$. In relation to this it is worth mentioning the work (Lattimore et al., 2011) which explores the possibility that a Bayesian predictor may fail to predict some subsequences.

5. Decision-Theoretic Interpretations of the Asymptotic Result

Classical decision theory is concerned with single-step games. Among its key results are the complete class and minimax theorems. The asymptotic formulation of the infinite-horizon problem considered here presents both differences and similarities which we attempt to summarize in this section. 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 that are common in decision theory. The terminology in this section is mainly after Ferguson (1967).

Predictors $\rho \in \mathcal{P}$ are called *strategies of the statistician*. The probability measures $\mu \in \mathcal{C}$ are now the basic *strategies of the opponent* (a.k.a. Nature), 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

$$E_{W(\mu)} \bar{L}(\mu, \rho), \tag{31}$$

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 $\nu = \int_{\mathcal{C}} \alpha dW(\alpha)$ is itself

a probability measure over the one-way infinite sequences, whereas a probability measure $W \in \mathcal{C}^*$ is a probability measure over $\mathcal{C}$.

5.1. Minimax

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

$$\bar{V}_{\mathcal{C}} := \inf_{\rho \in \mathcal{P}} \sup_{W \in \mathcal{C}^*} E_{W(\mu)} \bar{L}(\mu, \rho). \tag{32}$$

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

$$\underline{V}_{\mathcal{C}} := \sup_{W \in \mathcal{C}^*} \inf_{\rho \in \mathcal{P}} E_{W(\mu)} \bar{L}(\mu, \rho). \tag{33}$$

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 $\bar{L}(\mu, \rho)$ 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 $\bar{L}(\mu, \rho)$ is not (lower) semicontinuous with respect to it. Some other (including non-topological) sufficient conditions are given by Sion (1958); LeCam (1955). Other related results for KL divergence (expected log loss) include (Ryabko, 1979; Gallager, 1976-1979; Haussler, 1997).

In our setup, it is easy to see that, for every $\mathcal{C}$,

$$\bar{V}_{\mathcal{C}} = V_{\mathcal{C}}$$

and so Corollary 2 holds for $\bar{V}_{\mathcal{C}}$. Thus, using decision-theoretic terminology, we can state the following.

Corollary 4 (partial minimax theorem) *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 distribution part.

5.2. Complete Class

In this section we shall see that Corollary 2 can be interpreted as a complete-class theorem for asymptotic average loss, as well as some principled differences between the cases $V_{\mathcal{C}} > 0$ and $V_{\mathcal{C}} = 0$.

For a set of probability measures (strategies of the opponent) $\mathcal{C}$, a predictor ρ_1 is said to be *as good as* a predictor ρ_2 if $\bar{L}(\mu, \rho_1) \leq \bar{L}(\mu, \rho_2)$ 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 $\bar{L}(\mu, \rho_1) < \bar{L}(\mu, \rho_2)$ 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, 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 $\inf_{\rho \in \mathcal{P}} E_{W(\mu)} \bar{L}(\mu, \rho)$. 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; so the use of this terminology is non-contradictory here.

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:

Corollary 5 (Complete class theorem) *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 Corollary 2. 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 1), and a probability measure μ such that $\bar{L}(\mu, \rho) > 0$. Then for a predictor $\rho' := 1/2(\rho + \mu)$ we have $\bar{L}(\mu, \rho') = 0$. Thus, ρ' is better than ρ: its loss is strictly smaller on one measure, μ, and is at least the same on all the rest of the measures in $\mathcal{C}$. Therefore, ρ is inadmissible. The statement about minimal essentially complete class is proven analogously: indeed, take any essentially complete class, D, and any predictor $\rho \in D$. Take then the predictor ρ' constructed as above. Since ρ' is better than ρ and D is essentially complete, there must be another predictor $\rho'' \in D$, such that ρ'' is as good as ρ'. Therefore, $D \setminus \{\rho\}$ is essentially complete and D is not minimal. The statements about the case $V_{\mathcal{C}} = 0$ are self-evident. ■

6. Non-Realizable Case: Regret Minimization, Suboptimality of Bayes

In the non-realizable case the situation is principally different: it may happen that every combination of distributions in the model is suboptimal — even asymptotically.

Theorem 6 *There exist a set $\mathcal{C}$ for which $U_{\mathcal{C}} = 0$ and this value is attainable, yet for some constant $c > 0$ every Bayesian predictor φ with a prior concentrated on $\mathcal{C}$ must have*

Intuitively, the reason why any Bayesian predictor does not work in the counterexample of the proof given below is as follows. The set $\mathcal{C}$ considered is so large that any Bayesian predictor has to attach an exponentially decreasing a-posteriori weight to each element in $\mathcal{C}$. At the same time, by construction, each measure in $\mathcal{C}$ already attaches too little weight to the part of the event space on which it is a good predictor. In other words, the likelihood of the observations with respect to each predictor in $\mathcal{C}$ is too small to allow for any added penalty. To combine predictors in $\mathcal{C}$ one has to *boost* the likelihood, rather than attach a penalty. While this result is stated for Bayesian predictors, from the argument above it is clear that the example used in the proof is applicable to any combination of predictors in $\mathcal{C}$ one might think of, including, for example, MDL (Rissanen, 1989) and expert-advice-style predictors (e.g., Cesa-Bianchi and Lugosi, 2006). Indeed, if one has to boost the likelihood for some classes of predictors, it clearly breaks the predictor for other classes. In other words, there is no way to combine the prediction of the experts, short of disregarding them and doing something else instead.

Remark 7 (Countable $\mathcal{C}$) *Note that any set $\mathcal{C}$ satisfying the theorem must necessarily be uncountable. Indeed, for any countable set $\mathcal{C} = (\mu_k)_{k\in\mathbb{N}}$, take the Bayesian predictor $\varphi := \sum_{k\in\mathbb{N}} w_k\mu_k$, where w_k can be, for example, $\frac{1}{k(k+1)}$. Then, for any ν and any n, from (2) we obtain*

$$L_n(\nu,\varphi) \le -\log w_k + L_n(\nu,\mu_k).$$

That is to say, the regret of φ with respect to any μ_k is a constant independent of n (though it does depend on k), and thus for every ν we have $\bar{R}^\nu(C,\varphi) = 0$. It is worth noting that the origins of the use of such countable mixtures for prediction trace back to Zvonkin and Levin (1970); Solomonoff (1978).

Proof [of Theorem 6.] Let the alphabet $\mathcal{X}$ be ternary $\mathcal{X} = \{0,1,2\}$. For $\alpha \in (0,1)$ denote $h(\alpha)$ the binary entropy $h(\alpha) := -\alpha\log\alpha - (1-\alpha)\log(1-\alpha)$. Fix an arbitrary $p \in (0,1/2)$ and let β_p be the Bernoulli i.i.d. measure (produces only 0s and 1s) with parameter p. Let S be the set of sequences in $\mathcal{X}^\infty$ that have no 2s and such that the frequency of 1 is close to p:

$$S := \left\{\mathbf{x}\in\mathcal{X}^\infty : x_i \ne 2\forall i, \text{ and } \left|\frac{1}{t}|\{i=1..t : x_i = 1\}| - p\right| \le f(t) \text{ from some } t \text{ on}\right\},$$

where $f(t) = \log t/\sqrt{t}$. Clearly, $\beta_p(S) = 1$.

Define the set D_S as the set of all Dirac probability measures concentrated on a sequence from S, that is $D_S := \{\nu_{\mathbf{x}} : \nu_{\mathbf{x}}(\mathbf{x}) = 1,\ \mathbf{x}\in S\}$. Moreover, for each $\mathbf{x}\in S$ define the probability measure $\mu_{\mathbf{x}}$ as follows: $\mu_{\mathbf{x}}(X_{n+1}|X_{1..n}) = p$ coincides with β_p (that is, 1 w.p. p and 0 w.p. $1-p$) if $X_{1..n} = x_{1..n}$, and outputs 2 w.p. 1 otherwise: $\mu_{\mathbf{x}}(2|X_{1..n}) = 1$ if $X_{1..n} \ne x_{1..n}$. That is, $\mu_{\mathbf{x}}$ behaves as β_p only on the sequence $\mathbf{x}$, and on all other sequences it just outputs 2 deterministically. This means, in particular, that many sequences have probability 0, and some probabilities above are defined conditionally on zero-probability events, but this is not a problem; see the remark in the end of the proof.

Finally, let $\mathcal{C} := \{\mu_{\mathbf{x}} : \mathbf{x} \in S\}$. Next we define the predictor ρ that predicts well all measures in $\mathcal{C}$. First, introduce the probability measure δ that is going to take care of all the measures that output 2 w.p.1 from some time on. For each $a \in \mathcal{X}^*$ let δ_a be the probability measure that is concentrated on the sequence that starts with a and then consists of all 2s. Define $\delta := \sum_{a\in\mathcal{X}^*} w_a\delta_a$, where w_a are arbitrary positive numbers that sum to 1. Let also the probability measure β' be i.i.d. uniform over $\mathcal{X}$. Finally, define

$$\rho := 1/3(\beta_p + \beta' + \delta). \tag{34}$$

Next, let us show that, for every ν, the measure ρ predicts ν as well as any measure in $\mathcal{C}$: its loss is an additive constant factor. In fact, it is enough to see this for all $\nu \in D_S$, and for all measures that output all 2s w.p.1 from some n on. For each ν in the latter set, from (34) the loss of ρ is upper-bounded by $\log 3 - \log w_a$, where w_a is the corresponding weight. This is a constant (does not depend on n). For the former set, again from the definition (34) for every $\nu_{\mathbf{x}} \in D_S$ we have (see also Remark 7)

$$L_n(\nu_{\mathbf{x}}, \rho) \le \log 3 + L_n(\nu_{\mathbf{x}}, \beta_p) = nh(p) + o(n),$$

while

$$\inf_{\mu\in C} L_n(\nu_{\mathbf{x}}, \mu) = L_n(\nu_{\mathbf{x}}, \mu_{\mathbf{x}}) = nh(p) + o(n).$$

Therefore, for all ν we have

$$R_n^\nu(C, \rho) = o(n) \text{ and } \bar{R}^\nu(C, \rho) = 0.$$

Thus, we have shown that for every $\nu \in S$ there is a reasonably good predictor in $\mathcal{C}$ (here "reasonably good" means that its loss is linearly far from that of random guessing), and, moreover, there is a predictor ρ whose asymptotic regret is zero with respect to $\mathcal{C}$.

Next we need to show that any Bayes predictor has $2nh(p) + o(n)$ loss on at least some measure, which is double that of ρ, and which can be as bad as random guessing (or worse; depending on p). We show something stronger: any Bayes predictor has asymptotic average loss of $2nh(p)$ *on average* over all measures in S. So there will be many measures on which it is bad, not just one.

Let φ be any Bayesian predictor with its prior concentrated on $\mathcal{C}$. Since $\mathcal{C}$ is parametrized by S, for any $x_{1..n} \in \mathcal{X}^n, n \in \mathbb{N}$ we can write $\varphi(x_{1..n}) = \int_S \mu_{\mathbf{y}}(x_{1..n})dW(\mathbf{y})$ where W is some probability measure over S (the prior). Moreover, using the notation $W(x_{1..k})$ for the W-measure of all sequences in S that start with $x_{1..k}$, from the definition of the measures $\mu_{\mathbf{x}}$, for every $\mathbf{x} \in S$ we have

$$\int_S \mu_{\mathbf{y}}(x_{1..n})dW(\mathbf{y}) = \int_{\mathbf{y}\in S: y_{1..n}=x_{1..n}} \beta_p(x_{1..n})dW(\mathbf{y}) = \beta_p(x_{1..n})W(x_{1..n}). \tag{35}$$

Consider the average

$$E_U \limsup \frac{1}{n} L_n(\nu_x, \varphi)dU(\mathbf{x}),$$

where the expectation is taken with respect to the probability measure U defined as the measure β_p restricted to S; in other words, U is approximately uniform over this set. Fix

any $\nu_{\mathbf{x}} \in S$. Observe that $L_n(\nu_{\mathbf{x}}, \varphi) = -\log \varphi(x_{1..n})$. For the asymptotic regret, we can assume w.l.o.g. that the loss $L_n(\nu_{\mathbf{x}}, \varphi)$ is upper-bounded, say, by $n \log |\mathcal{X}|$ at least from some n on (for otherwise the statement already holds for φ). This allows us to use Fatou's lemma to bound

$$E_U \limsup \frac{1}{n} L_n(\nu_{\mathbf{x}}, \varphi) \geq \limsup \frac{1}{n} E_U L_n(\nu_{\mathbf{x}}, \varphi) = \limsup -\frac{1}{n} E_U \log \varphi(\mathbf{x})$$
$$= \limsup -\frac{1}{n} E_U \log \beta_p(x_{1..n}) W(x_{1..n}), \quad (36)$$

where in the last equality we used (35). Moreover,

$$- E_U \log \beta_p(x_{1..n}) W(x_{1..n})$$
$$= -E_U \log \beta_p(x_{1..n}) + E_U \log \frac{U(x_{1..n})}{W(x_{1..n})} - E_U \log U(x_{1..n}) \geq 2h(p)n + o(n), \quad (37)$$

where in the inequality we have used the fact that KL divergence is non-negative and the definition of U (that is, that $U = \beta_p|_S$). From this and (36) we obtain the statement of the theorem.

Finally, we remark that all the considered probability measures can be made non-zero everywhere by simply combining them with the uniform i.i.d. over $\mathcal{X}$ measure β', that is, taking for each measure ν the combination $\frac{1}{2}(\nu + \beta')$. This way all losses up to time n become bounded by $n \log |\mathcal{X}| + 1$, but the result still holds with a different constant. ■

7. Conclusion and Future Work

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 with 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 sequence, and to try to construct a predictor that minimizes the regret with respect to every distribution in the model class, on every sequence. Thus, one would be treating the model class as a set of experts. 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 sequence). Having done this, one may safely assume that the probability measure that generates the data belongs to the model class. 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. Indeed, 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

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, the choice of alternative approaches is very limited, but it would be interesting to see which predictors constructed for more specific settings can be generalized to arbitrary model classes, and whether the corresponding result holds for them.

Another interesting open 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 a discussion and some results on this property.

Some other interesting open questions are the decision-theoretic ones mentioned in Section 5; specifically, those concerning the minimax theorem and the existence of maximally spread distributions over $\mathcal{C}$. It would also be interesting to calculate the value $V_{\mathcal{C}}$ for different classes of distributions, similar to what is done for the i.i.d. example in Section 4.

Finally, several questions remain concerning the bounds presented in Section 3. The first question is how sharp is the upper bound in Theorem 1. So far, the lower bound (Theorem 3) 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 necessary, then the constant in front of the log becomes important, in particular 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.

References

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

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

P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *Annals of Statistics*, 14(1):1–26, 1986.

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

D. A.Freedman. On the asymptotic behavior of Bayes estimates in the discrete case I. *The Annals of Mathematical Statistics*, pages 1386–1403, 1963.

D. A. Freedman. On the asymptotic behavior of Bayes estimates in the discrete case II. *The Annals of Mathematical Statistics*, 36(2):454–456, 1965.

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

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

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

A. Gyorgy, T. Linder, and G. 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.

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

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

T. Lattimore, M. Hutter, and V. Gavane. Universal prediction of selected bits. In *Algorithmic Learning Theory*, pages 262–276. Springer, 2011.

L. 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. doi: 10.1109/18.556107.

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

J. J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific Publ. Co., 1989.

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

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

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

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

D. 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.

D. Ryabko. Universality of Bayesian mixture predictors. In *Proceedings of the 28th International Conference on Algorithmic Learning Theory (ALT'17)*, volume 76 of *PMLR*, pages 57–71, Kyoto, Japan, 2017. JMLR.

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

M. 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.

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

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

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

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