Predicting the outcomes of every process for which an asymptotically accurate stationary predictor exists is impossible

Daniil Ryabko
INRIA Lille, France
daniil@ryabko.net

Boris Ryabko
Institute of Computational Technologies,
SB RAS, Russia
boris@ryabko.net

***Abstract*—The problem of prediction consists in forecasting the conditional distribution of the next outcome given the past. Assume that the source generating the data is such that there is a stationary predictor whose error converges to zero (in a certain sense). The question is whether there is a universal predictor for all such sources, that is, a predictor whose error goes to zero if any of the sources that have this property is chosen to generate the data. This question is answered in the negative, contrasting a number of previously established positive results concerning related but smaller sets of processes.**

I. Introduction

The basic problem is predicting the conditional probability distribution $\hat{\mu}(\cdot|x_1,\dots,x_n)$ over the next outcome x_{n+1} given a sequence of observations $x_1,\dots,x_n$ generated by an unknown time-series distribution μ. Since $\hat{\mu}$ gives a conditional distribution for every $x_1,\dots,x_n$ it defines itself a time-series distribution. Thus, the source of data and the predictor are objects of the same kind. Traditionally, one assumes x_i to be independent and identically distributed, or that μ belongs to one of the well-studied parametric families. However, in applications involving hard-to-model data sources such as stock market, human-written texts or biological sources, it is often assumed, instead, that μ belongs to some large (nonparametric) family of time-series distributions. Examples of such families are the set of all finite-memory distributions or the set of all stationary distributions. The hope is not that the unknown data source under study actually belongs to such a family – for example, that a human-written text obeys the finite-memory assumption or that the stock market is stationary – but, rather, that the considered family of sources is good enough for the forecasting task at hand. Such a "hope," however, remains informal, since the theoretical results concern the setting when the unknown source belongs to the family.

Here we consider a formalization of the "good enough for prediction" setting proposed in [1]. Specifically, we are asking whether a predictor can be constructed which is asymptotically consistent (prediction error goes to 0) on any source for which a consistent predictor exists in a given family. Thus, given a set $\mathcal{S}$ of distributions, we consider the set $\mathcal{S}^+ := \{$*of all distributions μ such that there exists a distribution $\nu \in \mathcal{S}$ such that the prediction error of ν on sequences generated by μ goes to zero*$\}$. We are asking whether there exists a predictor that is consistent on all distributions in $\mathcal{S}^+$. In this work the family $\mathcal{S}$ in question is the set of all stationary ergodic distributions, and the question is answered in the negative. This negative result is rather tight; in particular, the same proof shows that the set S can be replaced by the set of all Hidden Markov chains with a countable set of states (maintaining the negative result), while a consistent predictor exists if we only consider Hidden Markov chains with a finite set of states.

Prior work. A consistent predictor for the set $\mathcal{S}$ of all finite-alphabet stationary distributions has been constructed in [2]. Here the prediction quality is with respect to Cesaro-averaged Kullback-Leibler (KL) divergence, which is required to converge to 0 either in expectation or with probability 1. The same work shows that an analogous result is impossible to obtain without Cesaro averaging (the latter negative result was obtained independently in the unpublished thesis of Bailey [3]). The positive result admits a number of generalizations and extensions, including those to continuous alphabets [4], [5], [6], [7], [8], [9].

Prediction with *expert advice* (see [10] for an overview) presents a different approach to the problem of prediction. Here one assumes that the data source to predict is an arbitrary deterministic sequence, and makes no further assumptions on it. The goal is also different: rather than trying to make the prediction error decrease to 0 (which is impossible in this setting), it is required to predict as well as any expert from a given set. An important difference is that in this setting one does not give probability forecasts of the next outcome but just deterministic predictions, and the quality is measured (according to some loss function) with respect to the prediction of each expert. The set of experts is usually small, most typically finite; the class of all i.i.d. predictors also has been considered [11]. While this approach is very close to the one taken here, it does not allow one to look at predictors (experts) and data source as objects of the same kind, thus making it difficult to formulate our question of interest.

A connection between the settings was made in the work [1], which formulates three problems. The first one is the classical

problem of constructing a predictor that is asymptotically consistent (its error goes to 0) if any process from an (arbitrary, given) set $\mathcal{S}$ is chosen to generate the data. The second is the one considered in this work: asymptotically consistent prediction of sequences generated by every source for which there is an asymptotically consistent predictor in a given set $\mathcal{S}$. The third setting removes the "asymptotically consistent" part: it requires constructing a predictor that predicts any source whatsoever as well as any predictor in a given set $\mathcal{S}$. Thus, the latter formulation is the worst-case analysis akin to expert advice (the only difference is that we still try to forecast probabilities, rather than individual outcomes). Here all predictors and sources are just time-series distributions. The three problems are naturally ordered in difficulty: if the set $\mathcal{S}$ is the same in all the three problems, then any solution to the third problem is a solution to the second, and any solution to the second is a solution to the first. For the set of all stationary processes, it is known since [2] that the first problem admits a solution. It is shown in [1] that the third problem (worst-case) does not, but the question of whether the second problem admits a solution for this set was left open; here we answer it in the negative.

II. Preliminaries

Let $\mathcal{X}$ be a finite set. Since we are after a negative result, selecting $\mathcal{X} := \{0,1\}$ is not a restriction, so we fix this choice. The notation $x_{1..n}$ is used for $x_1,\dots,x_n$. We consider time-series distributions, that is, probability measures on $\Omega := (\mathcal{X}^\infty, \mathcal{B})$ where $\mathcal{B}$ is the sigma-field generated by the cylinder sets $[x_{1..n}]$, $x_i \in \mathcal{X}, n \in \mathbb{N}$ and $[x_{1..n}]$ is the set of all infinite sequences that start with $x_{1..n}$. We use $\mathbb{E}_\mu$ for the expectation with respect to a measure μ.

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

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

In words, we take the expected (over data) cumulative (over time) KL divergence between μ- and ρ-conditional (on the past data) probability distributions of the next outcome. Define also

$$d(\mu,\rho) := \liminf_{n\to\infty} \frac{1}{n} d_n(\mu,\rho).$$

We say that ρ *predicts* μ (in expected average KL divergence) if $d(\mu,\rho)=0$. It is easy to see that

$$d_n(\mu,\rho) = \mathbb{E}_\mu \log \frac{\mu(x_{1..n})}{\rho(x_{1..n})},$$

which makes expected average KL divergence a convenient measure of prediction quality to study.

Let the set $\mathcal{P}$ be the set of all time-series distributions over Ω. A distribution $\rho\in\mathcal{P}$ is *stationary* if for every $i,j\in\mathbb{N}$ and every $A\in\mathcal{X}^j$, we have

$$\rho(X_{1..j}=A) = \rho(X_{i..i+j-1}=A).$$

A stationary distribution ρ is called *ergodic* if for all $n\in\mathbb{N}, A\in\mathcal{X}^n$ with probability 1 we have $\lim_{n\to\infty}\nu(X_{1..n},A) = \rho(A)$, where $\nu(X_{1..n},A)$ stands for the frequency of occurrence of the word A in $X_{1..n}$. (The latter definition can be shown to be equivalent to the usual one formulatedf in terms of shift-invariant sets [12].)

III. Main result

Denote $\mathcal{S}\subset\mathcal{P}$ the set of all stationary ergodic time-series distributions. Define

$$\mathcal{S}^+ := \{\mu\in\mathcal{P} : \exists \nu\in\mathcal{S}\ d(\nu,\mu)=0\}.$$

Theorem 1. *For any predictor $\rho\in\mathcal{P}$ there is a measure $\mu\in\mathcal{S}^+$ such that $d(\mu,\rho)\geq 1$.*

Proof: We will show that the set $\mathcal{S}^+$ includes the set $\mathcal{D}$ of all Dirac measures, that is, of all measures concentrated on one deterministic sequence. The statement of the theorem follows directly from this, since for any ρ one can find a sequence $x_1,\dots,x_n,\dots\in\mathcal{X}^\infty$ such that $\rho(x_n|x_{1..n-1})\leq 1/2$ for all $n\in\mathbb{N}$.

To show that $\mathcal{D}\subset\mathcal{S}^+$, we will construct, for any given sequence $x := x_1,\dots,x_n,\dots\in\mathcal{X}^\infty$, a measure μ_x such that $d(\delta_x,\mu_x)=0$ where δ_x is the Dirac measure concentrated on x. These measures are constructed as functions of a stationary Markov chain with a countably infinite set of states. The construction is based on the one used in [2] (see also [7]).

The Markov chain M has the set $\mathbb{N}$ of states. From each state j it transits to the state $j+1$ with probability $p_j := j^2/(j+1)^2$ and to the state 1 with the remaining probability, $1-p_j$. Thus, M spends most of the time around the state 1, but takes rather long runs towards outer states: long, since p_j tends to 1 rather fast. We need to show that it does not "run away" too much; more precisely, we need to show M has a stationary distribution. For this, it is enough to show that the state 1 is positive recurrent (see, e.g., [13, Chapter VIII] for the definitions and facts about Markov chains used here). This can be verified directly as follows. Denote $f_{11}^{(n)}$ the probability that starting from the state 1 the chain returns to the state 1 for the first time in exactly n steps. We have

$$f_{11}^{(n)} = (1-p_n)\prod_{i=1}^{n-1} p_i = \left(1-\frac{n^2}{(n+1)^2}\right)\frac{1}{n^2}.$$

To show that the state 1 is positive recurrent we need $\left(\sum_{n=0}^\infty n f_{11}^{(n)}\right)^{-1} > 0$. Indeed, $n f_{11}^{(n)} < 3/n^2$ which is summable. It follows that M has a stationary distribution, which we call π.

For a given sequence $x := x_1,\dots,x_n,\dots\in A^\infty$, the measure μ_x is constructed as a function g_x of the chain M taken with its stationary distribution as the initial one. We define $g_x(j) = x_j$ for all $j\in\mathbb{N}$. Since M is stationary, so

is μ_x. It remains to show that $d(\delta_x, \mu_x) = 0$. Indeed, we have

$$d_n(\delta_x, \mu_x) = -\log \mu_x(x_1, \dots, x_n) \le -\log \left(\pi_1 \prod_{j=1}^{n} p_j \right) = -\log \pi_1 + 2\log(n+1) = o(n).$$

■

A. Other sets of measures to predict and tightness of the result

One can ask how "tight" is the negative result presented, or, in other words, whether the set $\mathcal{S}$ was too general a point of departure in the first place.

To answer this question, first note that, as mentioned before, the work [2] shows (by an explicit construction) that there is a universal predictor for the set $\mathcal{S}$ (of stationary ergodic distributions) itself, that is, there exist a measure ρ such that $d(\mu, \rho) = 0$ for any $\mu \in \mathcal{S}$.

Next, from the proof of Theorem 1 one can see that it is possible to replace the set $\mathcal{S}$ in its formulation with the set of all hidden Markov chains with a countably infinite set of states. The latter set is in fact much smaller than the set $\mathcal{S}$. Indeed, $\mathcal{S}$ can be considered as the set of all stationary hidden Markov processes with an uncountably infinite (specifically, $\mathcal{X}^\infty$) set of states, giving the "much smaller" comparison above a precise set-theoretic meaning.

Passing to positive results for the problem of prediction considered, for the set $\mathcal{M}$ of all finite-memory processes, [1, Theorem 15] shows that there is a universal predictor for the set $\mathcal{M}^+$. Moreover, it is easy to extend the proof of the latter result to all hidden Markov processes with finitely many states. Thus, it is possible to predict all measures that are predicted by a hidden Markov chain with finitely many states, but not with a countably infinite set of states, making the negative result rather tight.

B. Other measures of prediction quality

So far, we have been measuring the quality of prediction in terms of expected average KL divergence. Measuring it differently would change both the set $\mathcal{S}^+$ and the requirement on the predictor that would have to predict all measures from this set. Thus, the result of Theorem 1 does not directly entail a similar statement about neither weaker nor stronger measures of prediction quality.

However, a quick look at the proof of Theorem 1 shows that the construction it employs is rather universal. Specifically, μ_x predicts x also almost surely rather than in expectation (simply because the sequence x is a deterministic sequence).

Moreover, the prediction error (of μ_x on x) convergence to 0 in just about any sense one can think of, for example, one can replace KL divergence with the absolute loss, squared loss, etc. This implies that Theorem 1 holds for these measures of prediction quality as well. Furthermore, this shows that the same result holds if we consider different notions of prediction on different sides of the question: asking whether it is possible to predict in (say) expected average KL divergence all measures that are predicted by some stationary ergodic measure when (say) the convergence has to be with probability 1 and there is no time-averaging.

Thus, the result (placed in the title) appears to be general and not an artefact of the measure of prediction quality considered.

Acknowledgments

Daniil Ryabko acknowledges the support of the French Ministry of Higher Education and Research, the Nord-Pas-de-Calais Regional Council and FEDER. Boris Ryabko acknowledges the support of the Russian Foundation for Basic Research, project no. 15-07-01851

REFERENCES

[1] D. Ryabko, "On the relation between realizable and non-realizable cases of the sequence prediction problem." *Journal of Machine Learning Research*, vol. 12, pp. 2161–2180, 2011.

[2] B. Ryabko, "Prediction of random sequences and universal coding," *Problems of Information Transmission*, vol. 24, pp. 87–96, 1988.

[3] D. H. Bailey, *Sequential schemes for classifying and predicting ergodic processes*. Department of Mathematics, Stanford University., 1976.

[4] P. Algoet, "Universal schemes for prediction, gambling and portfolio selection," *The Annals of Probability*, vol. 20, no. 2, pp. 901–941, 1992.

[5] G. Morvai, S. Yakowitz, and L. Gyorfi, "Nonparametric inference for ergodic, stationary time series," *Ann. Statist*, vol. 24, no. 1, pp. 370–379, 1996.

[6] G. Morvai, S. Yakowitz, and P. Algoet, "Weakly convergent nonparametric forecasting of stationary time series," *Information Theory, IEEE Transactions on*, vol. 43, no. 2, pp. 483 –498, Mar. 1997.

[7] L. Gyorfi, G. Morvai, and S. Yakowitz, "Limits to consistent on-line forecasting for ergodic time series," *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 886–892, 1998.

[8] B. Ryabko, "Compression-based methods for nonparametric prediction and estimation of some characteristics of time series," *IEEE Transactions on Information Theory*, vol. 55, pp. 4309–4315, 2009.

[9] ——, "Applications of universal source coding to statistical analysis of time series," *Selected Topics in Information and Coding Theory, World Scientific Publishing*, pp. 289–338, 2010.

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

[11] Y. Freund, "Predicting a binary sequence almost as well as the optimal biased coin," *Information and Computation*, vol. 182, no. 2, pp. 73–94, 2003.

[12] R. Gray, *Probability, Random Processes, and Ergodic Properties*. Springer Verlag, 1988.

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