This is a draft of the paper presented at the 2019 IEEE International Symposium of Information Theory. Please cite as:

D. Kalociński and T. Steifer, "An Almost Perfectly Predictable Process with No Optimal Predictor", 2019 IEEE International Symposium on Information Theory (ISIT), Paris, France, 2019, pp. 2504-2508, doi: 10.1109/ISIT.2019.8849587.

An Almost Perfectly Predictable Process with No Optimal Predictor

Dariusz Kalociński
Institute of Philosophy
University of Warsaw
ul. Krakowskie Przedmieście 3
00-927 Warsaw, Poland
Email: dkalocinski@uw.edu.pl*

Tomasz Steifer
Institute of Fundamental Technological Research
Polish Academy of Sciences
ul. Pawińskiego 5B
02-106 Warsaw, Poland
Email: tsteifer@ippt.pan.pl

Abstract

A novel kind of a negative result is presented for the problem of computable prediction. A non-stationary binary stochastic process is constructed for which almost surely no effective method of prediction achieves the infimum of prediction errors defined as the normalized Hamming distance between the sequence of predictions and the realization of the process. Yet it is shown that this process may be effectively predicted almost surely up to an arbitrarily small error since the infimum of prediction errors is zero.

1 Introduction

The question of prediction or forecasting is centuries old. In information-theoretic setting it was studied in particular by Shannon in [1]. In this context, 'prediction' may be explicated in several slightly different ways. Some

*Please use: dariusz.kalocinski@gmail.com

authors by 'prediction' mean experimental estimation of unknown conditional probabilities [2]. In this approach, one may study asymptotic convergence of the estimator, the vanishing of quadratic difference between estimates and true probabilities, or use some other familiar measures such as Kullback-Leibler divergence. Universal estimators exist for sources such as finite Markov chains or stationary ergodic processes [2].

A different meaning of 'prediction' may be found in [3]. Here, at the beginning of each period a predictor is expected to choose an action from a space of possible actions. At each moment, a finite number of discrete observations is available. Using information about these, the predictor makes a choice of some particular action such as a guess of a most probable outcome of the next bit of some binary process. Then, the action is evaluated according to some appropriate loss function. The performance is measured by comparing predictor's choice with the outcome of the particular realization of the process in the next period. In particular, this may simply mean that the predictor guesses the next value in the realization. In such scenario an elegant loss function is given by a simple ratio of correct to all predictions. Such notion was also studied in theoretical computer science—specifically, in the context of algorithmic randomness and stochasticity [4, 5].

In probabilistic setting, a considerable attention was given to universal procedures for prediction of wide classes of stochastic processes. In particular, Algoet [3] studied prediction of stationary ergodic processes. The minimum conditional expected loss is bounded from below by an optimal strategy based on a priori known probabilities. Since the probabilities are rarely known in real applications, one may sought universal methods that are asymptotically optimal for a reasonably large class of processes. Indeed, such universal predictors exist for stationary ergodic processes.

The scope of this paper is limited to prediction of binary processes understood as an effort to guess the next bit of the process. Of course, what constitutes a reasonable predictor is subject to discussion. For practical reasons one may choose to limit the scope of interest to some class of predictors, e.g., finite-state predictors [6, 7]. Here, it is assumed that a minimal constraint for predictors is computability as understood in theoretical computer science and, in particular, in computability theory (for an introduction to computability theory, see e.g., [8, 9]). Computability, although restrictive, is a relatively weak constraint on effectiveness.

Following the requirement of computability, we assume that an effective predictor should be implementable in a form of a computer program in some programming language such as C or Python. Note that some of the known universal prediction schemes are not effective in this sense. For example, the

predictor described in [10] requires access to an external source of randomness and it is not known whether this randomization can be avoided.

In general, not every computable function is total—it may not halt for some inputs. It is not clear how to measure performance of a predictor that is sometimes undefined. Therefore, we further limit our attention only to those predictors that are both computable and total. Such predictors will be called proper predictors. For proper predictors, prediction is defined for every (finite) sequence of values. This is satisfied, for example, by deterministic predictors with finite memory but the class of proper predictors is not limited to these.

A simple but important observation can be made about the cardinality of the set of proper predictors. Every proper predictor may be associated with a program which computes it. A computer program is a finite string of letters over a finite alphabet. Therefore, it may be uniquely coded by a natural number. Consequently, the set of all predictors is countable. This allows to use various diagonalization arguments to construct binary sequences or stochastic processes which are in some way out of the reach of every proper predictor. Similar arguments may be made for different problems such as effective measure estimation. For an example of such construction see, e.g., [11].

In the sequel of the paper, we will present an example of a non-stationary binary process of some paradoxical property. It will be proven that almost surely no proper predictor is optimal for this process. In other words, for every proper predictor there exists another proper predictor which is almost surely better than the former one. In particular, every predictor is asymptotically worse than the improper predictor based on the perfect a priori information about the underlying measure. That being said, the process in question is far from being unpredictable. Indeed, we will show that proper prediction suffices to predict the process, almost surely, with an arbitrarily small error. The present theorem is an extension of a similar result proven for deterministic case in [12].

2 Prequisities

Symbol λ denotes the empty string. For non-empty strings $\sigma \in \{0,1\}^*$, we will write $\sigma = \sigma_1 \dots \sigma_n = \sigma_1^n$, where $\sigma_i \in \{0,1\}$ are individual bits. Similarly, for infinite sequences $x \in \{0,1\}^\omega$, we will write $x = x_1 x_2 ...$, where $x_i \in \{0,1\}$. Using these notations, we introduce the following two definitions.

Definition 1 (predictors)**.** *A proper predictor is a total computable function*

$f : \{0,1\}^* \to \{0,1\}$. *An improper predictor is a total function* $f : \{0,1\}^* \to \{0,1\}$ *which is not computable.*

Definition 2 (prediction errors)**.** *Let* f *be an (im)proper predictor and let* $\sigma \in \{0,1\}^*$ *be a non-empty string,* $\sigma \neq \lambda$*. The prediction error of* f *on* σ *is defined as*

$$\varsigma(f,\sigma) := \frac{\#\{1 \leq i \leq |\sigma| : \sigma_i \neq f(\sigma_1^{i-1})\}}{|\sigma|} \tag{1}$$

For an infinite sequence $x \in \{0,1\}^\omega$*, the prediction errors of* f *on* x *are defined as*

$$\varsigma_+(f,x) := \limsup_{n\to\infty} \varsigma(f,x_1^n), \tag{2}$$

$$\varsigma_-(f,x) := \liminf_{n\to\infty} \varsigma(f,x_1^n), \tag{3}$$

whereas we write $\varsigma(f,x) := \varsigma_+(f,x) = \varsigma_-(f,x)$ *if the later two limits are equal.*

3 Statement of the result

Theorem 1. *There exists a binary stochastic process* $X = X_1X_2...$ *such that for every proper predictor* f *there is a proper predictor* g *such that*

$$\varsigma_-(g,X) < \varsigma_-(f,X) \text{ almost surely.} \tag{4}$$

Moreover, for every $\epsilon > 0$ *there exists a proper predictor* f *such that*

$$\varsigma_+(f,X) < \epsilon \text{ almost surely.} \tag{5}$$

The proof of Theorem 1 uses a version of the diagonal argument and relies on the fact that the set of total computable functions (in particular, the set of all proper predictors) is countable. Thus, we can use some enumeration of all proper predictors in our construction. The process will be constructed inductively. At each step, a different predictor will be under consideration, selected according to a predefined function $p : \mathbb{N} \to \mathbb{N}$, assigning $p(n)$-th predictor (from a chosen non-effective enumeration of all proper predictors) to the n-th bit of the process. The corresponding measure, for bits assigned to a given predictor, will be defined in a way that ensures that the predictor almost surely fails at predicting correctly almost all of those bits. Specifically, the function p assigning predictors to the bits of the process will be defined

$h_1 \longrightarrow$	·	X_1	·	X_3	·	X_5	·	X_7	·	X_9	·	X_{11}	·	X_{13}	·	...
$h_2 \longrightarrow$	·	·	X_2	·	·	·	X_6	·	·	·	X_{10}	·	·	·	·	...
$h_3 \longrightarrow$	·	·	·	·	X_4	·	·	·	·	·	·	·	X_{12}	·	·	...
⋮																

Table 1: The assigment of random bits X_i of the process to predictors h_k

so as to guarantee that the first predictor in the enumeration will be wrong in the limit almost surely at least once for every two bits, the second predictor will be wrong at least once for every four bits, and so on.

Crucially, the function p which assigns bits of the process to predictors has the following property: given a predictor f, there exists an effective procedure which recognizes the bits corresponding to the predictor—namely, those bits where we can expect f to be wrong for the most of the time. So, we can use this to correct the prediction by acting inversely to f when needed. In this way we can guarantee the existence of a better proper predictor.

On the other hand, a well kwown fact in computability theory is that the enumeration of all (and only) proper predictors cannot be effective, that is, it cannot be realized by any computable function. (To see this, assume for a contradiction that $f_1, f_2, \ldots$ is a computable enumeration of all proper predictors and define a proper predictor f such that $f(w_n) = 1 - f_n(w_n)$, where w_n is the n-th binary string in the bounded lexicographical order. Clearly, f is a proper predictor but it is not in the list $f_1, f_2, \ldots.$) Since there is no computable enumeration of all proper predictors, we cannot use the correction procedure to all proper predictors. This ensures that no predictor is optimal and perfect (i.e., no predictor achieves the prediction error equal 0).

But assigning predictors to infinitely many bits in adequately sparse way and setting appropriate probabilities, we can also ensure that the process may be predicted with an arbitrarily small error. This will follow from the fact that you can take an arbitrary finite number of proper predictors, use them as subprocedures and act inversely.

4 Proof of Theorem 1

We begin with this auxiliary observation.

Lemma 1. *For every even number $i > 0$ there exist unique integers $k \geq 1$*

and $n \geq 0$ *such that*

$$i = (1/2 + n)2^k. \tag{6}$$

Moreover, i *is odd if and only if* $k = 1$.

Let $h_1, h_2, \ldots$ be an (uncomputable) listing of all proper predictors. We start by defining an assignment $p : \mathbb{N} \to \mathbb{N}$, so that the predictor $h_{p(i)}$ will be assigned to the i-th random bit of the process X. By Lemma 1, for each natural number i there are unique integers $k \geq 1$ and $n \geq 0$ such that

$$i = (1/2 + n)2^k.$$

We set $p(i) := k$. It is easy to verify that for each k, the predictor h_k is assigned to a bit of the process X once per 2^k bits. We will use this observation later on. Table 1 shows in a visual way how the bits of process X are assigned to the predictors.

Now, we proceed to construct the probability distribution of process X inductively. Firstly, set

$$P(X_1 = h_{p(1)}(\lambda)) = 1.$$

Subsequently, we set iteratively for $i \in \mathbb{N}$ that

$$P(X_{i+1} = h_{p(i+1)}(\sigma) | X_1^i = \sigma) = \frac{1}{(i+1)^2}.$$

In other words, the probability that $h_{p(k)}$ makes a correct prediction on k-th bit is equal to $1/k^2$. Now, let $i_1, i_2, \ldots$ be all natural numbers such that for some m,

$$m = p(i_1) = p(i_2) = \ldots$$

Observe that

$$\sum_{i \in \{i_1, i_2 \ldots\}} P(h_m \text{ correctly predicts } i_1\text{-th bit}) < \infty.$$

Hence, by the Borel-Cantelli lemma, predictor h_m is correct on finitely many bits with indices from $\{i_1, i_2, \ldots\}$ almost surely.

Observe that $i_{n+1} - i_n = 2^m$. Consequently, we have

$$\varsigma_-(h_m) > 2^{-m} \text{ almost surely.}$$

We will construct a predictor g which is almost surely better than h_m. We know on which bits predictor h_m is correct only finitely many times and

we know that these bits are placed once per every 2^m bits. Since we can effectively compute the indices of these bits, we can use that information to alter the prediction when it is desirable. Let for all $\sigma \in \{0,1\}^*$

$$g(\sigma) = h(\sigma) \text{ iff } p(|\sigma|) \neq m.$$

Note that on bits with indexes from $\{i_1, i_2, \ldots\}$ predictor g is wrong only finitely many times while predictor h_m is correct only finitely many times (almost surely). Moreover, on the rest of the bits these predictors always agree. Hence

$$\varsigma_-(g, X) = \varsigma_-(h_m, X) - 2^{-m} \text{ almost surely.}$$

Since m was chosen arbitrarily, this completes the proof of the first part of Theorem 1.

To demonstrate the second part of Theorem 1, we will show that for every $\epsilon > 0$ there is a predictor f such that

$$\varsigma_+(f, X) < \epsilon \text{ almost surely.}$$

Fix an $\epsilon > 0$. Let $k > 0$ be the smallest number such that

$$2^{-k} = 1 - \sum_{i=1}^{k} \frac{1}{2^i} < \epsilon.$$

We will construct a predictor f such that

$$\varsigma_+(f, X) \leq 2^{-k} \text{ almost surely.}$$

We already know that, in the limit, the first predictor h_1 is almost surely wrong at least on the half of the bits, the second predictor h_2 is almost surely wrong at least once for every four bits, and so on. We can compute the indexes on which this happens. We will require that f makes a different prediction than $h_1, \ldots, h_k$ on the corresponding bits. This will guarantee that almost surely f is asymptotically correct at least on fraction

$$1 - 2^{-k} = \sum_{i=1}^{k} \frac{1}{2^i}$$

of the bits. To be precise, we set

$$f(\sigma) = \begin{cases} 1 - h_{p(|\sigma|)}(\sigma) & p(|\sigma|) \leq k, \\ 0 & \text{otherwise.} \end{cases}$$

Since k is finite and we can compute $p(|\sigma|)$ for every σ, f is a proper predictor. Consequently,

$$\varsigma_+(f, X) \leq 2^{-k} < \epsilon \text{ almost surely.}$$

5 Conclusions

In this paper, we have constructed a binary stochastic process with two important properties. On the one hand, the process may be effectively predicted very well, up to an arbitrarily small prediction error. Indeed, for a given positive threshold ϵ, there exists a proper (i.e., total computable) predictor with a prediction error almost surely smaller than ϵ. On the other hand, no such predictor is optimal. Indeed, for any proper predictor, there is another proper predictor with an almost surely smaller prediction error. Consequently, the constructed process escapes any effective method of prediction while being arbitrarily well predictable.

The assumption of effectiveness plays an important role in our construction. Firstly, proper predictors form a countable set which allows us to use diagonalization within a discrete infinite binary process. Secondly, proper predictors, being computer programs, are closed on subroutinization and complementation. Subroutinization means that any finite number of proper predictors can be made into a new proper predictor that inherits the behaviour of its subroutines. Closure on complementation guarantees that flipping the behaviour of a proper predictor yields another proper predictor (in this context, it is perhaps worth noting that closure on complementation is not common to all computing devices—consider, for example, nondeterministic pushdown automata [8]). These two features are used when it is demonstrated that one can predict the process up to an arbitrary small error. Thirdly, no enumeration of proper predictors is effective which essentially leads to non-optimality of all predictors for the constructed process.

It seems, however, that any model of computation that shares the above properties should be prone to a construction similar to the one presented here. For example, it seems that restricting predictors to deterministic finite-state automata should yield a similar result. Another interesting point is that the process constructed in this way reflects an inherent constraint of a given model of sequential prediction in general. Although the constructed process is not optimally predictable by any predictor from the predefined set of admissible prediction strategies, there is a predictor outside this set that can predict the process with zero error. However, such a predictor requires more computational power. In the present paper, the stochastic process in question is random and not effectively computable, while the predictors are deterministic. Hence, it is reasonable to ask for similar results for nondeterministic prediction, for example predictors giving answers according to some (computable) measure. Again, given sufficiently strong model of prediction, similar diagonalization argument may be applied.

Acknowledgment

The authors would like to thank both referees for valuable comments. Moreover, the authors are indebted to Łukasz Dębowski for sharing his expertise, useful discussions, proofreading and making the paper more readable to information theory community.

References

[1] C. E. Shannon, "Prediction and entropy of printed English," *Bell System Technical Journal*, vol. 30, no. 1, pp. 50–64, 1951.

[2] B. Y. Ryabko, "Prediction of random sequences and universal coding," *Problems of Information Transmission*, vol. 24, no. 2, pp. 87–96, 1988.

[3] P. H. Algoet, "The strong law of large numbers for sequential decisions under uncertainty," *IEEE Transactions on Information Theory*, vol. 40, no. 3, pp. 609–633, 1994.

[4] K.-I. Ko, "On the notion of infinite pseudorandom sequences," *Theoretical Computer Science*, vol. 48, pp. 9–33, 1986.

[5] K. Ambos-Spies, E. Mayordomo, Y. Wang, and X. Zheng, "Resource-bounded balanced genericity, stochasticity and weak randomness," in *Annual Symposium on Theoretical Aspects of Computer Science*. Springer, 1996, pp. 61–74.

[6] M. Feder, N. Merhav, and M. Gutman, "Universal prediction of individual sequences," *IEEE Transactions on Information Theory*, vol. 38, no. 4, pp. 1258–1270, 1992.

[7] E. Meron and M. Feder, "Finite-memory universal prediction of individual sequences," *IEEE Transactions on Information Theory*, vol. 50, no. 7, pp. 1506–1523, 2004.

[8] J. Hopcroft and J. Ullman, *Introduction to automata theory, languages, and computation*. Addison-Wesley Publishing Company, Reading, MA, 1979.

[9] R. I. Soare, *Turing computability: Theory and applications*. Springer, 2016.

[10] L. Gyorfi, G. Lugosi, and G. Morvai, "A simple randomized algorithm for sequential prediction of ergodic time series," *IEEE Transactions on Information Theory*, vol. 45, no. 7, pp. 2642–2650, 1999.

[11] H. Takahashi, "Computational limits to nonparametric estimation for ergodic processes," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 6995–6999, 2011.

[12] D. Kalociński and T. Steifer, "On unstable and unoptimal prediction," *Mathematical Logic Quarterly*, forthcoming.