Universal probability-free prediction

Vladimir Vovk and Dusko Pavlovic

April 5, 2017

Abstract

We construct universal prediction systems in the spirit of Popper's falsifiability and Kolmogorov complexity and randomness. These prediction systems do not depend on any statistical assumptions (but under the IID assumption they dominate, to within the usual accuracy, conformal prediction). Our constructions give rise to a theory of algorithmic complexity and randomness of time containing analogues of several notions and results of the classical theory of Kolmogorov complexity and randomness.

The conference version of this paper has been published in the Proceedings of COPA 2016. The journal version is to appear in the Special Issue of the *Annals of Mathematics and Artificial Intelligence* devoted to COPA 2016. The version at http://alrw.net (Working Paper 14) is updated most often.

1 Introduction

In this paper we consider the problem of predicting the labels, assumed to be binary, of a sequence of objects. This is an online version of the standard problem of binary classification. Namely, we will be interested in infinite sequences of observations

$$\omega = (z_1, z_2, \ldots) = ((x_1, y_1), (x_2, y_2), \ldots) \in (\mathbf{X} \times 2)^\infty$$

(also called *infinite data sequences*), where $\mathbf{X}$ is an *object space* and $2 := \{0, 1\}$ is the *label space*. For simplicity, we will assume that the object space $\mathbf{X}$ is a given finite set of, say, binary strings (the intuition being that finite objects can always be encoded as binary strings). The elements 1 and 0 of the label space are often interpreted as "true" and "false".

Finite sequences $\sigma \in (\mathbf{X} \times 2)^*$ of observations will be called *finite data sequences*. If σ_1, σ_2 are two finite data sequences, their concatenation will be denoted (σ_1, σ_2); σ_2 is also allowed to be an element of $\mathbf{X} \times 2$. A standard partial order on $(\mathbf{X} \times 2)^*$ is defined as follows: $\sigma_1 \sqsubseteq \sigma_2$ means that σ_1 is a prefix of σ_2; $\sigma_1 \sqsubset \sigma_2$ means that $\sigma_1 \sqsubseteq \sigma_2$ and $\sigma_1 \neq \sigma_2$. The smallest element in this order (the empty data sequence) is denoted $\square$. We say that finite data sequences σ_1 and σ_2 are *comparable* if $\sigma_1 \sqsubseteq \sigma_2$ or $\sigma_2 \sqsubseteq \sigma_1$.

We use the notation $\mathbb{N} := \{1, 2, \ldots\}$ for the set of positive integers and $\mathbb{N}_0 := \{0, 1, 2, \ldots\}$ for the set of nonnegative integers. The *length* of a finite data sequence σ is the number $l \in \mathbb{N}_0$ such that $\sigma \in (\mathbf{X} \times 2)^l$. If $\omega \in (\mathbf{X} \times 2)^\infty$ and $l \in \mathbb{N}_0$, $\omega^l \in (\mathbf{X} \times 2)^l$ is the prefix of ω of length l.

We will also use the notation $\sigma_1 \sqsubseteq \sigma_2$ and $|\sigma|$ for finite binary sequences $\sigma_1, \sigma_2, \sigma \in 2^*$.

A *situation* is a concatenation $(\sigma, x) \in (\mathbf{X} \times 2)^* \times \mathbf{X}$ of a finite data sequence σ and an object x; our task in the situation (σ, x) is to be able to predict the label of the new object x given the sequence σ of labelled objects. Given a situation $s = (\sigma, x)$ and a label $y \in 2$, we let (s, y) stand for the finite data sequence $(\sigma, (x, y))$, which is the concatenation of s and y.

Our notation for binary logarithm will be log.

The contents of this paper

This paper is, to some degree, a result of our attempts to understand the philosophical problem of prediction. It has two components, philosophical and mathematical, and the latter is more or less independent of the former. The main goal of the remainder of this section is to provide a road map for our mathematical readers who do not share our philosophy of science, so that the latter does not get in their way. The four key mathematical concepts introduced in this paper are:

- universal prediction system (Sections 2–6),
- time complexity (Sections 6–7),
- *a priori* time semimeasure (Section 8),
- time randomness (Section 9).

In the remaining sections we will explore (rather superficially) various connections between these key concepts.

We start from a toy formalization of the philosophical notion of a law of nature in Section 2 and the most basic way of using laws of nature for prediction in Section 3. The notion of a strong prediction system introduced in Section 3 has only philosophical interest in this paper, and this section can be safely skipped by our mathematical readers. The notion of a weak prediction system introduced in the following Section 4 is more convenient from the mathematical point of view since there exists a universal weak prediction system, as shown in Section 5. Section 6 introduces the notion of complexity for weak prediction systems and uses it to strengthen the property of universality of the universal prediction system.

The reader who is not interested in prediction can start from the mathematical notion of laws of nature in Section 2 and the definition of their complexity in the second part of Section 6, which will prepare her to reading Section 7 about time complexity (apart from the theorem describing connections with universal prediction).

The definition of the *a priori* time semimeasure in Section 8 is self-contained and does not depend on the previous sections. This section contains simple connections between the *a priori* time semimeasure and time complexities.

Section 9 devoted to time randomness is the nexus of this paper. Time randomness is defined in terms of time complexity (another natural definition would be in terms of *a priori* time semimeasure) and serves as the basis for prediction under conditions of considerable noise (including connections with conformal prediction).

The last two sections, 10 and 11, explain connections of the key mathematical concepts of this paper with the theories of conformal prediction and Kolmogorov complexity, respectively.

2 Laws of nature as prediction systems

> Not for nothing do we call the laws of nature "laws": the more they prohibit, the more they say.
>
> *The Logic of Scientific Discovery*
> Karl Popper

According to Popper's [7] view of the philosophy of science, scientific laws of nature should be falsifiable: some finite sequences of observations should disagree with such a law, and if so, we should be able to detect the disagreement. (Popper often preferred to talk about scientific theories or statements instead of laws of nature. He did not discuss the computational details of detecting disagreement; for him, it was just something that we see straight away.) The empirical content of a law of nature is the set of its potential falsifiers ([7], Sections 31 and 35). We start from formalizing this notion in our toy setting, interpreting the requirement that we should be able to detect falsification as that we should be able to detect it eventually.

Formally, we define a *law of nature* L to be a recursively enumerable prefix-free subset of $(\mathbf{X} \times 2)^*$ (where *prefix-free* means that $\sigma_1 \notin L$ whenever $\sigma_2 \in L$ and $\sigma_1 \sqsubset \sigma_2$). Intuitively, these are the potential falsifiers, i.e., sequences of observations prohibited by the law of nature. The requirement of being recursively enumerable is implicit in the notion of a falsifier, and the requirement of being prefix-free reflects the fact that extensions of prohibited sequences of observations are automatically prohibited and there is no need to mention them in the definition (see, however, Remark 2 below). It is convenient to allow the vacuous law of nature $\emptyset$.

A law of nature L gives rise to a prediction system: in a situation $s = (\sigma, x)$ it predicts that the label $y \in 2$ of the new object x will be an element of

$$\Pi_L(s) := \{y \in 2 \mid (s, y) \notin L\}. \tag{1}$$

There are three possibilities in each situation s:

- The law of nature makes a prediction, either 0 or 1, in situation s when the prediction set (1) is of size 1, $|\Pi_L(s)| = 1$.
- The prediction set is empty, $|\Pi_L(s)| = 0$, which means that the law of nature is about to be falsified (and we can even say that it has been falsified already).
- The law of nature refrains from making a prediction when $|\Pi_L(s)| = 2$. This can happen in two cases:
  - the law of nature was falsified in past: $\sigma' \in L$ for some $\sigma' \sqsubseteq \sigma$;
  - the law of nature has not been falsified as yet and allows $(\sigma, (x, y))$ for both $y = 0$ and $y = 1$.

Remark. The counterpart of our notion of a law of nature in probability theory is that of a stopping time.

Remark. Our definition of a law of nature is the one that appears to us to lead to the simplest and richest theory, but there are several viable alternatives. Let us say that a subset L of $(\mathbf{X} \times 2)^*$ is an *upset* if the conjunction of $\sigma_1 \in L$ and $\sigma_1 \sqsubseteq \sigma_2$ implies $\sigma_2 \in L$. It is clear that the definition of laws of nature as upsets L whose *frontier*

$$\{\sigma \in L \mid \forall \sigma' \sqsubset \sigma : \sigma' \notin L\}$$

is recursively enumerable is completely equivalent to ours; talking about frontiers of upsets rather than upsets is a matter of taste. We can both narrow down and widen up this definition in a natural way. The most restrictive definition discussed in this remark identifies a law of nature with a computable upset L in $(\mathbf{X} \times 2)^*$ that is *strongly co-continuable*, in the sense of satisfying

$$\forall \sigma \notin L \, \forall x \in \mathbf{X} \, \exists y \in 2 : (\sigma, (x, y)) \notin L$$

(which is equivalent to the prediction set (1) never being empty unless the law has been falsified). A slightly less restrictive definition is to identify a law of nature with a computable upset that is *weakly co-continuable*, in the sense of satisfying

$$\forall \sigma \notin L \, \exists (x, y) \in (\mathbf{X} \times 2) : (\sigma, (x, y)) \notin L.$$

(Notice that a subset of $(\mathbf{X} \times 2)^*$ is a weakly co-continuable upset if and only if it can be represented as the set of all finite data sequences all of whose infinite continuations are elements of a given open subset of $(\mathbf{X} \times 2)^\infty$; this gives a natural one-to-one correspondence between the weakly co-continuable upsets and the open sets in $(\mathbf{X} \times 2)^\infty$.) The main reasons we do not impose either of the conditions of co-continuability are that we opt for simpler definitions and that empty prediction sets are common in conformal prediction. A serious disadvantage of definitions involving the requirement of computability is that, in non-trivial cases, they do not allow constructing universal objects (such as universal prediction systems in Section 5 below). Dropping the two conditions of co-continuability and relaxing computability to recursive enumerability

of the frontier, we obtain our definition. Further relaxing the requirement of computability, we can define a law of nature as a recursively enumerable upset. This is a natural definition that fits our intuition behind laws of nature even better than our official definition does; it also allows us to define universal predictions systems (as in Theorems 1 and 3 below). However, we do not know how to connect it with natural counterparts of the standard notions of Kolmogorov complexity, *a priori* semimeasure, and randomness (as we do for our definition in Sections 7–9). And our official definition still covers the practically important case of computable upsets.

3 Strong prediction systems

The notion of a law of nature is static; experience tells us that laws of nature eventually fail and are replaced by other laws. Popper represented his picture of this process by formulas ("evolutionary schemas") similar to

$$\mathrm{PS}_1 \to \mathrm{TT}_1 \to \mathrm{EE}_1 \to \mathrm{PS}_2 \to \cdots \tag{2}$$

(introduced in his 1965 talk on which [8], Chapter 6, is based and also discussed in several other places in [8] and [9]). In response to a problem situation PS, scientists create a tentative theory TT and then subject it to attempts at error elimination EE, whose success leads to a new problem situation PS, after which scientists come up with a new tentative theory TT, etc. In our toy version of this process, tentative theories are laws of nature, problem situations are situations in which our current law of nature becomes falsified, and there are no active attempts at error elimination (so that error elimination simply consists in waiting until the current law of nature becomes falsified).

If L and L' are laws of nature, we define $L \sqsubset L'$ to mean that for any $\sigma' \in L'$ there exists $\sigma \in L$ such that $\sigma \sqsubset \sigma'$. To formalize the philosophical picture (2), we define a *strong prediction system* $\mathcal{L}$ to be a nested sequence $L_1 \sqsubset L_2 \sqsubset \cdots$ of laws of nature $L_1, L_2, \ldots$ that are jointly recursively enumerable, in the sense of the set $\{(\sigma, n) \in (\mathbf{X} \times 2)^* \times \mathbb{N} \mid \sigma \in L_n\}$ being recursively enumerable.

The interpretation of a strong prediction system $\mathcal{L} = (L_1, L_2, \ldots)$ is that L_1 is the initial law of nature used for predicting the labels of new objects until it is falsified; as soon as it is falsified we start looking for and then using for prediction the following law of nature L_2 until it is falsified in its turn, etc. Therefore, the prediction set in a situation $s = (\sigma, x)$ is natural to define as the set

$$\Pi_{\mathcal{L}}(s) := \{y \in 2 \mid (s, y) \notin \cup_{n=1}^{\infty} L_n\}. \tag{3}$$

As before, it is possible that $\Pi_{\mathcal{L}}(s) = \emptyset$.

Fix a situation $s = (\sigma, x) \in (\mathbf{X} \times 2)^* \times \mathbf{X}$. Let $n = n(s)$ be the largest integer such that σ has a prefix in L_n. It is possible that $n = 0$ (when s does not have such prefixes), but if $n \geq 1$, s will also have prefixes in $L_{n-1}, \ldots, L_1$, by the definition of a strong prediction system. Then L_{n+1} will be the current law of nature; all earlier laws, $L_n, L_{n-1}, \ldots, L_1$, have been falsified. The prediction (3)

in situation s is then interpreted as the set of all labels y that are not prohibited by the current law L_{n+1}.

In the spirit of the theory of Kolmogorov complexity, we would like to have a universal prediction system. However, we are not aware of any useful notion of a universal strong prediction system. Therefore, in the next section we will introduce a wider notion of a prediction system that does not have this disadvantage.

4 Weak prediction systems

A *weak prediction system* $\mathcal{L}$ is defined to be a sequence (not required to be nested in any sense) $L_1, L_2, \ldots$ of laws of nature $L_n \subseteq (\mathbf{X} \times 2)^*$ that are jointly recursively enumerable.

Remark. Popper's evolutionary schema (2) was the simplest one that he considered; his more complicated ones, such as

$$\begin{array}{rl} & \nearrow \mathrm{TT_a} \to \mathrm{EE_a} \to \mathrm{PS_{2a}} \to \cdots \\ \mathrm{PS_1} \to & \mathrm{TT_b} \to \mathrm{EE_b} \to \mathrm{PS_{2b}} \to \cdots \\ & \searrow \mathrm{TT_c} \to \mathrm{EE_c} \to \mathrm{PS_{2c}} \to \cdots \end{array}$$

(cf. [8], pp. 243 and 287), give rise to weak rather than strong prediction systems.

In the rest of this paper we will omit "weak" in "weak prediction system". The most basic way of using a prediction system $\mathcal{L}$ for making a prediction in situation $s = (\sigma, x)$ is as follows. Decide on the maximum number N of errors you are willing to make. Ignore all L_n apart from $L_1, \ldots, L_N$ in $\mathcal{L}$, so that the prediction set in situation s is

$$\Pi^N_{\mathcal{L}}(s) := \{y \in 2 \mid \forall n \in \{1, \ldots, N\} : (s, y) \notin L_n\}\,.$$

Notice that this way we are guaranteed to make at most N mistakes: making a mistake eliminates at least one law in the list of unfalsified laws among $\{L_1, \ldots, L_N\}$.

Similarly to the theory of conformal prediction (see, e.g., [14]), another way of packaging $\mathcal{L}$'s prediction in situation s is, instead of choosing the threshold (or *level*) N in advance, to allow the user to apply her own threshold: in a situation s, for each $y \in 2$ report the attained level

$$\pi^s_{\mathcal{L}}(y) := \min\{n \in \mathbb{N} \mid (s, y) \in L_n\} \in \mathbb{N} \cup \{\infty\} \tag{4}$$

(with $\min \emptyset := \infty$). The user whose threshold is N will then consider $y \in 2$ with $\pi^s_{\mathcal{L}}(y) \le N$ as prohibited in s. Notice that the function (4) is upper semicomputable (for a fixed $\mathcal{L}$).

The strength of a prediction system $\mathcal{L} = (L_1, L_2, \ldots)$ at level $N \in \mathbb{N}$ is determined by its *N-part*

$$\mathcal{L}_{\le N} := \bigcup_{n=1}^{N} L_n. \tag{5}$$

At level N, the prediction system $\mathcal{L}$ prohibits $y \in 2$ as continuation of a situation s if and only if $(s, y) \in \mathcal{L}_{\le N}$.

We will also use the limit

$$\mathcal{L}_{<\infty} := \bigcup_{n=1}^{\infty} L_n \tag{6}$$

of (5). Notice that $\mathcal{L}_{<\infty}$ uniquely determines $\mathcal{L}$ if $\mathcal{L}$ is a strong prediction system, but the analogous statement for weak prediction systems is false.

Remark. In motivating our definitions we have referred to views expressed in Karl Popper's writings. Similar views have been held by many other philosophers. Popper himself regarded his evolutionary schemas as improvements and rationalizations of the Hegelian dialectical schema. Charles Peirce's views were particularly close to Popper's. He was as emphatic as Popper in insisting on the importance of falsification of laws of nature (as he said, "the scientific spirit requires a man to be at all times ready to dump his whole cartload of beliefs, the moment experience is against them" [6, pp. 46–47]). His version of Popper's evolutionary schema (2) is

$$\text{belief} - \text{surprise} - \text{doubt} - \text{inquiry} - \text{belief}$$

(as presented by Misak in [5, p. 11]).

5 Universal prediction

> There is the logical disjunction: Either an intrinsically improbable event will occur, or, the prediction will [...] be verified.
>
> *Statistical Methods and Scientific Inference*
> Ronald Fisher

The following theorem says that there exists a universal prediction system, in the sense that it is stronger than any other prediction system if we ignore a multiplicative increase in the number of errors made.

Theorem 1. *There is a* universal *prediction system $\mathcal{U}$, in the sense that for any prediction system $\mathcal{L}$ there exists a constant $c > 0$ such that, for any N,*

$$\mathcal{L}_{\le N} \subseteq \mathcal{U}_{\le cN}. \tag{7}$$

Proof. Let $\mathcal{L}^1, \mathcal{L}^2, \dots$ be a recursive enumeration of all prediction systems; their component laws of nature will be denoted $(L_1^k, L_2^k, \dots) := \mathcal{L}^k$. (Formally, we require the set $\{(\sigma, n, k) \in (\mathbf{X} \times 2)^* \times \mathbb{N}^2 \mid \sigma \in L_n^k\}$ to be recursively enumerable and the sequence $\mathcal{L}^1, \mathcal{L}^2, \dots$ to contain all prediction systems.) For each $n \in \mathbb{N}$, define the nth component U_n of $\mathcal{U} = (U_1, U_2, \dots)$ as follows. Let the binary representation of n be

$$\left(a, 0, 1^{k-1}\right) = (a, 0, 1, \dots, 1),$$

where a is a binary string (starting from 1) and the number of 1s in the $1, \dots, 1$ is $k - 1 \in \mathbb{N}_0$ (this sentence is the definition of $a = a(n)$ and $k = k(n)$ in terms of n). If the binary representation of n does not contain any 0s, a and k are undefined, and we set $U_n := \emptyset$. Otherwise, set

$$U_n := L_A^k,$$

where $A \in \mathbb{N}$ is the number whose binary representation is a. In other words, $\mathcal{U}$ consists of the components of $\mathcal{L}^k$, $k \in \mathbb{N}$; namely, L_1^k is placed in $\mathcal{U}$ as $U_{3 \times 2^{k-1}-1}$ and then $L_2^k, L_3^k, \dots$ are placed at intervals of 2^k:

$$U_{3 \times 2^{k-1}-1+2^k(n-1)} = L_n^k, \quad n = 1, 2, \dots.$$

It is easy to see that

$$\mathcal{L}_{\le N}^k \subseteq \mathcal{U}_{\le 3 \times 2^{k-1}-1+2^k(N-1)}, \tag{8}$$

which is stronger than (7). □

Let us fix a universal prediction system $\mathcal{U}$. We can equivalently rewrite (7) as the inclusion between the extreme terms of

$$\Pi_{\mathcal{U}}^{cN}(s) = \{y \in 2 \mid (s, y) \notin \mathcal{U}_{\le cN}\} \subseteq \{y \in 2 \mid (s, y) \notin \mathcal{L}_{\le N}\} = \Pi_{\mathcal{L}}^{N}(s), \tag{9}$$

for all situations s. Intuitively, (9) says that the prediction sets output by the universal prediction system are at least as precise as the prediction sets output by any other prediction system $\mathcal{L}$ if we ignore a constant factor in specifying the level N.

In terms of the attained level (4), Theorem 1 says that, as a function of s and y, $\pi_{\mathcal{U}}^s(y)$ does not exceed $\pi_{\mathcal{L}}^s(y)$ to within a constant factor. Indeed, assuming that $c \in \mathbb{N}$ in (7),

$$\begin{aligned}
\pi_{\mathcal{L}}^s(y) &= \min\{n \in \mathbb{N} \mid (s, y) \in L_n\} = \max\{N \in \mathbb{N}_0 \mid (s, y) \notin \mathcal{L}_{\le N}\} + 1 \\
&\ge \max\{N \in \mathbb{N}_0 \mid (s, y) \notin \mathcal{U}_{\le cN}\} + 1 \\
&= \frac{1}{c} \max\{N' \in c\mathbb{N}_0 \mid (s, y) \notin \mathcal{U}_{\le N'}\} + 1 \\
&\ge \frac{1}{c} \max\{N' \in \mathbb{N}_0 \mid (s, y) \notin \mathcal{U}_{\le N'}\} \\
&= \frac{1}{c} \min\{n \in \mathbb{N} \mid (s, y) \in U_n\} - \frac{1}{c} = \frac{1}{c} \pi_{\mathcal{U}}^s(y) - \frac{1}{c},
\end{aligned}$$

which implies

$$\pi^s_{\mathcal{L}}(y) \geq \frac{1}{2c}\pi^s_{\mathcal{U}}(y) \tag{10}$$

when $\pi^s_{\mathcal{U}}(y) \geq 2$; and when $\pi^s_{\mathcal{U}}(y) = 1$, (10) follows from $\pi^s_{\mathcal{L}}(y) \geq 1$.

If we are in a situation $s = (\sigma, x)$ and one of the two $\pi^s_{\mathcal{U}}(y)$ (corresponding to $y = 0$ or $y = 1$) is a small number, we can predict the other label: e.g., if $\pi^s_{\mathcal{U}}(0)$ is small, we can predict that the label of x is 1, and we then have Fisher's disjunction: either our prediction is correct, or a rare event has occurred.

6 Complexity of prediction systems and laws of nature

In this section we will see how the constant c in Theorem 1 depends on the prediction system $\mathcal{L}$. The dependence will be in terms of the algorithmic complexity of $\mathcal{L}$, which we will now define.

A *description language for prediction systems* is a function F mapping 2^* to the set of all prediction systems such that the set

$$\{(d, \sigma, n) \in 2^* \times (\mathbf{X} \times 2)^* \times \mathbb{N} \mid \sigma \in F_n(d)\}$$

is recursively enumerable, where $F_n(d)$ is the nth law of nature in $F(d)$, i.e., $F_n(d) := L_n$ when $F(d) = (L_1, L_2, \ldots)$. Notice that the domain of F is 2^* rather than a subset of 2^*, which is unusual for the theory of algorithmic complexity. The *effective domain* $\mathrm{dom}(F)$ of a description language F for prediction systems is

$$\mathrm{dom}(F) := \{d \mid F(d) \neq (\emptyset, \emptyset, \ldots)\}. \tag{11}$$

A *prefix-free description language for prediction systems* is a description language F for prediction systems such that $\mathrm{dom}(F)$ is prefix-free.

The *complexity* of a prediction system $\mathcal{L}$ with respect to a description language F for prediction systems is defined by

$$C_F(\mathcal{L}) := \min\{|d| \mid \mathcal{L} = F(d)\},$$

$|d|$ standing for the length of d.

Theorem 2. *There is a description language U for prediction systems that is universal in the sense that for any description language F for prediction systems there exists a constant c such that, for any prediction system $\mathcal{L}$,*

$$C_U(\mathcal{L}) \leq C_F(\mathcal{L}) + c. \tag{12}$$

There is a prefix-free description language U' for prediction systems that is universal in the sense that for any prefix-free description language F for prediction systems there exists a constant c such that, for any prediction system $\mathcal{L}$,

$$C_{U'}(\mathcal{L}) \leq C_F(\mathcal{L}) + c.$$

Proof. We will use the same (very standard) argument as in Theorem 1 and will only prove (12). Let F^k, $k = 1, 2, \ldots$ be a recursive enumeration of the description languages for prediction systems (meaning that the set

$$\left\{(d, \sigma, n, k) \in 2^* \times (\mathbf{X} \times 2)^* \times \mathbb{N}^2 \mid \sigma \in F^k_n(d)\right\}$$

is recursively enumerable and that each description language for prediction systems belongs to the sequence $F^1, F^2, \ldots$). Let $1^k 0 d$ serve as a description of $F^k(d)$ under U (where 1^k stands for the binary sequence $(1, \ldots, 1)$ consisting of k 1s). □

Let us fix a universal description language U for prediction systems, call $C_U(\mathcal{L})$ the *plain complexity* of $\mathcal{L}$, and abbreviate $C_U(\mathcal{L})$ to $C(\mathcal{L})$. Analogously, we fix a universal prefix-free description language U', call $C_{U'}(\mathcal{L})$ the *prefix complexity* of $\mathcal{L}$, and abbreviate $C_{U'}(\mathcal{L})$ to $K(\mathcal{L})$.

The following theorem makes (7) uniform in $\mathcal{L}$ showing how c depends on $\mathcal{L}$.

Theorem 3. *There is a constant $c > 0$ such that, for any prediction system $\mathcal{L}$ and any $N \in \mathbb{N}$, the universal prediction system $\mathcal{U}$ satisfies*

$$\mathcal{L}_{\le N} \subseteq \mathcal{U}_{\le c 2^{K(\mathcal{L})} N}. \tag{13}$$

Proof. Define a prediction system $\mathcal{V}$ as the sequence $(V_1, V_2, \ldots)$ of laws of nature such that $V_n := U'_{n'}(d)$, where U' is the universal prefix-free description language for prediction systems, and $n' \in \mathbb{N}$ and $d \in 2^*$ are defined given n as follows:

- d is the suffix (if it exists) of the binary representation of n such that $\overleftarrow{d}$ belongs to $\mathrm{dom}(U')$ (where $\overleftarrow{d}$ is the mirror image of d: $\left|\overleftarrow{d}\right| = |d|$ and the bits of $\overleftarrow{d}$ are the same as the bits of d but written in the opposite order);
- the binary representation of n' is the prefix (if non-empty) of the binary representation of n left after removing its suffix d.

It is clear that such n' and d are unique when they exist; and when they do not exist, set $V_n := \emptyset$. Then the modification

$$U'_n(d) \subseteq \mathcal{V}_{\le n 2^{|d|} + 2^{|d|} - 1}$$

of (8) implies, for any prediction system $\mathcal{L}$,

$$\mathcal{L}_n \subseteq \mathcal{V}_{\le n 2^{K(\mathcal{L})} + 2^{K(\mathcal{L})} - 1}$$

(take as d the shortest description of $\mathcal{L}$ under U'). This implies that (13) holds for some prediction system $\mathcal{V}$ in place of $\mathcal{U}$, which, when combined with the statement of Theorem 1, implies that (13) holds for our chosen universal prediction system $\mathcal{U}$. □

Specializing the notions of plain and prefix complexity for a prediction system to prediction systems of type $\mathcal{L} = (L, L, \ldots)$, we obtain the notions of plain and prefix complexity for a law of nature:

$$C(L) := C((L, L, \ldots)),$$
$$K(L) := K((L, L, \ldots)).$$

However, since the notion of algorithmic complexity of a law of nature will be used in the next section as a basis for defining the complexity of time, we will also spell out the simpler direct definition.

A *description language for laws of nature* is a function F mapping 2^* to the set of prefix-free subsets of $(\mathbf{X} \times 2)^*$ such that the set

$$\{(d, \sigma) \in 2^* \times (\mathbf{X} \times 2)^* \mid \sigma \in F(d)\}$$

is recursively enumerable. We will usually omit "for laws of nature". Notice that, for any description language F and any *description* $d \in 2^*$, $F(d)$ is a law of nature (formally, we use "description" to mean elements of 2^*; informally, descriptions serve as arguments for description languages). The *effective domain* $\mathrm{dom}(F)$ of a description language F is

$$\mathrm{dom}(F) := \{d \mid F(d) \neq \emptyset\}.$$

A *prefix-free description language* is a description language F such that $\mathrm{dom}(F)$ is prefix-free.

The *complexity* of a law of nature L with respect to a description language F is defined by

$$C_F(L) := \min \{|d| \mid L = F(d)\}.$$

The analogue of Theorem 2 continues to hold for laws of nature; we fix a universal description language U, call $C_U(L)$ the *plain complexity* of L, and abbreviate $C_U(L)$ to $C(L)$; and we fix a universal prefix-free description language U', call $C_{U'}(L)$ the *prefix complexity* of L, and abbreviate $C_{U'}(L)$ to $K(L)$.

This is a corollary of Theorem 3 for laws of nature:

Corollary 1. *There is a constant $c > 0$ such that, for any law of nature L, the universal prediction system $\mathcal{U}$ satisfies*

$$L \subseteq \mathcal{U}_{\leq c 2^{K(L)}}. \tag{14}$$

Proof. We again regard laws of nature L as a special case of prediction systems identifying L with $\mathcal{L} := (L, L, \ldots)$. It remains to apply Theorem 3 to $\mathcal{L}$ setting $N := 1$. □

A simple counting argument shows that the dependence of the right-hand side of (13) on the complexity of $\mathcal{L}$ is approximately correct and cannot be significantly improved (if the difference between plain and prefix complexities is

ignored). To state this argument in its strongest form, we will introduce a new piece of notation: for each infinite data sequence $\omega \in (\mathbf{X} \times 2)^\infty$,

$$\Sigma(\omega) := \left\{\omega^l \mid l \in \mathbb{N}_0\right\}$$

is the set of all finite prefixes of ω. Theorem 3 says that there is a constant $c > 0$ such that, for any $K, N \in \mathbb{N}$, any infinite data sequence ω, and any prediction system $\mathcal{L}$ satisfying $K(\mathcal{L}) \le K$,

$$\mathcal{L}_{\le N} \cap \Sigma(\omega) \subseteq \mathcal{U}_{\le c2^K N} \cap \Sigma(\omega). \tag{15}$$

The inclusion in (15) compares the predictive powers of $\mathcal{L}$ and $\mathcal{U}$ only along the infinite data sequence ω.

Theorem 4. *There is a constant $c > 0$ such that, for any $K, N \in \mathbb{N}$ and any infinite data sequence ω, there exists a prediction system $\mathcal{L}$ satisfying $C(\mathcal{L}) \le K$ and*

$$\mathcal{L}_{\le N} \cap \Sigma(\omega) \not\subseteq \mathcal{U}_{\le c2^K N} \cap \Sigma(\omega). \tag{16}$$

Proof. Let $\mathcal{L}^k$, $k \in \mathbb{N}$, be the strong prediction system such that $\mathcal{L}^k_{<\infty}$ (defined by (6)) consists of finite data sequences whose length is divisible by 2^{k-1} but not divisible by 2^k (what is essential is that different $\mathcal{L}^k$ make errors on disjoint sets of finite data sequences). Take any $K, N \in \mathbb{N}$ and any $\omega \in (\mathbf{X} \times 2)^\infty$. Set $K' := K - a$ for some constant $a \in \mathbb{N}$, to be chosen later. The set $\mathcal{U}_{\le 2^{K'} N} \cap \Sigma(\omega)$ contains at most $2^{K'} N$ elements; therefore, (16) will be satisfied for $c := 2^{-a}$, for some $\mathcal{L} := \mathcal{L}^k$ and $k \le 2^{K'} + 1$. It remains to notice that $C(\mathcal{L}^k) \le K' + O(1) \le K$ provided a is sufficiently large. □

We have the following corollary of Theorem 4 for laws of nature showing the tightness (to within the difference between C and K) of Corollary 1.

Corollary 2. *There is a constant $c > 0$ such that, for any $K \in \mathbb{N}$ and any infinite data sequence ω, there exists a law of nature L satisfying $C(L) \le K$ and*

$$L \cap \Sigma(\omega) \not\subseteq \mathcal{U}_{\le c2^K} \cap \Sigma(\omega).$$

Proof. Specialize Theorem 4 to the case $N := 1$ and define L to be the first element of the prediction system $\mathcal{L}$. The additive constant implicit in the definition of the plain complexity $C(L)$ can be incorporated into the constant c, as we did in the proof of Theorem 4. □

Analogously to (7) and (9), we can rewrite (13) and (14) as

$$\Pi^{c2^{K(\mathcal{L})}N}_{\mathcal{U}}(s) \subseteq \Pi^N_{\mathcal{L}}(s) \tag{17}$$

and

$$\Pi^{c2^{K(L)}}_{\mathcal{U}}(s) \subseteq \Pi_L(s), \tag{18}$$

respectively, for all situations s; (17) and (18) indicate the dependence of the constant factor in (9) on $\mathcal{L}$.

Remark ([13])**.** This is a natural modification of our definition of prefix-free description languages: a description language F for laws of nature is *prefix-correct* if, for all $d_1, d_2 \in 2^*$,

$$d_1 \sqsubseteq d_2 \Longrightarrow F(d_1) \subseteq F(d_2).$$

There is a universal prefix-correct description language U'' in the sense that $C_{U''} \leq C_F + O(1)$ for any prefix-correct description language F. Let us fix such a U'' and call $K'(F) := C_{U''}(F)$ the *intermediate complexity* of F.

7 Time complexity of finite data sequences

The *plain time complexity* and *prefix time complexity* of a finite data sequence σ are defined by

$$\mathbf{C}(\sigma) := \min_{L \ni \sigma} C(L), \tag{19}$$

$$\mathbf{K}(\sigma) := \min_{L \ni \sigma} K(L), \tag{20}$$

respectively, where L ranges over the laws of nature. (We will explain the terminology later in this section.) We have to modify the notation C and K slightly since we would like to be able to use the standard notation $C(\sigma)$ and $K(\sigma)$ for the Kolmogorov complexity (plain and prefix) of σ; we will also use $C(n)$ and $K(n)$ to denote the Kolmogorov complexity (plain or prefix) of an integer n.

The following simple result is useful for discussing the interpretation of $\mathbf{C}(\sigma)$ and $\mathbf{K}(\sigma)$.

Theorem 5. *For any finite data sequence σ,*

$$\mathbf{C}(\sigma) \leq C(|\sigma|) + O(1), \tag{21}$$

$$\mathbf{K}(\sigma) \leq K(|\sigma|) + O(1). \tag{22}$$

Proof. If F is any description language (prefix-free or not) for nonnegative integers, we can define a description language F' for laws of nature by setting

$$F'(d) := \begin{cases} (\mathbf{X} \times 2)^{F(d)} & \text{if } F(d) \text{ is defined} \\ \emptyset & \text{if not.} \end{cases}$$

Since

$$C_{F'}\left((\mathbf{X} \times 2)^{|\sigma|}\right) = C_F(|\sigma|),$$

(21) follows by setting $F := U$ and (22) follows by setting $F := U'$. □

Theorem 5 gives a trivial bound on the time complexity of σ: it is the complexity of the length of σ (i.e., of the time of the last observation in σ

assuming that the observations are taken at times $1, 2, \ldots$). We can say that both $\mathbf{C}(\sigma)$ and $\mathbf{K}(\sigma)$ measure the complexity of the time of the last observation in σ when we are given the observations themselves as an oracle (with the observations disclosed sequentially, so that we can't just count them). As we will see later (see Theorems 6 and 10 below), these measures of complexity can be used to determine whether being in the situation of having just observed the last observation in σ is a rare event[1]. For the purpose of prediction, having such a measure of complexity is important since our prediction system can be forgiven for giving a wrong prediction when a rare event happens (cf. the epigraph about "Fisher's disjunction" to Section 5).

Remark. The length of a finite data sequence σ can be interpreted as the physical time of the last observation in σ. In probability theory, physical time is often changed; e.g., it can be replaced by intrinsic time reflecting the intensity at which various events happen (in a probability-free setting, this was done in, e.g., [15], where physical time was replaced by quadratic variation). The stopping times (see Remark 2) corresponding to physical time consist of all finite data sequences of the same length. For the more general notion of time, we can regard the last observations in the finite data sequences in an arbitrary stopping time (law of nature) as happening at the same moment in time. This is another justification for calling (19)–(20) the time complexity of σ.

Remark. In the usual jargon of Kolmogorov complexity, we can say that the complexity (either plain or prefix) of σ is the minimal complexity (of the same kind) of a binary program that enumerates some prefix-free set containing σ.

The following theorem describes a connection with the universal prediction system; remember that log is binary logarithm.

Theorem 6. *There is a constant $c > 0$ such that, for all N,*

$$\{\sigma \mid \mathbf{C}(\sigma) \le \log N - c\} \subseteq \mathcal{U}_{\le N} \subseteq \{\sigma \mid \mathbf{C}(\sigma) \le \log N + c\}. \tag{23}$$

Proof. To check the left-hand inclusion in (23), it suffices to define a prediction system $\mathcal{L}$ such that, for all finite data sequences σ, $\sigma \in \mathcal{L}_{\le 2^{k+1}}$ where $k := \mathbf{C}(\sigma)$. Let U be the universal description language for laws of nature: $C_U = \mathbf{C}$. We can set $\mathcal{L} := (L_1, L_2, \ldots)$, where L_n is defined to be $F(d)$ for d obtained from the binary representation of n by removing the leading 1.

To check the right-hand inclusion in (23), it suffices to define a description language F for laws of nature such that $C_F(\sigma) \le \log N$ whenever $\sigma \in \mathcal{U}_{\le N}$, for any N. Define $F(d)$, where $d \in 2^*$, as U_n, where n is the natural number whose binary representation is 1 followed by d. If $\sigma \in \mathcal{U}_{\le N}$, σ will belong to a law of nature whose description is of length at most $\lfloor \log N \rfloor$, which completes the proof of this inclusion. □

[1]For the reader familiar with Shafer's ([11], Section 1.7) distinction between Humean and Moivrean events, we are talking about events of the former kind.

We can interpret (23) by saying that $\mathcal{U}_{\leq N}$ coincides with $\{\sigma \mid \mathbf{C}(\sigma) \leq \log N\}$ if we are allowed to vary the threshold $\log N$ by adding a constant (positive or negative); this qualification is natural as time complexity is defined only to within an additive constant.

The next result gives an even simpler connection.

Theorem 7. *When $s \in (\mathbf{X} \times 2)^* \times \mathbf{X}$ ranges over the situations and $y \in 2$ over the labels,*

$$\log \pi^s_{\mathcal{U}}(y) = \mathbf{C}((s, y)) + O(1).$$

Proof. This follows immediately from (23):

$$\begin{aligned}\log \pi^s_{\mathcal{U}}(y) &= \min\{\log n \mid (s, y) \in U_n\} = \min\{\log N \mid (s, y) \in \mathcal{U}_{\leq N}\} \\ &= \min\{\log N \mid \mathbf{C}((s, y)) \leq \log N\} + O(1) = \mathbf{C}((s, y)) + O(1),\end{aligned}$$

where n and N range over $\mathbb{N}$. □

And the following theorem gives obvious connections between the two complexities.

Theorem 8.

$$\begin{aligned}\mathbf{C}(\sigma) &\leq \mathbf{K}(\sigma) + O(1) \\ \mathbf{K}(\sigma) &\leq \mathbf{C}(\sigma) + 2 \log \mathbf{C}(\sigma) + O(1).\end{aligned}$$

Proof. The first inequality follows from the fact that a prefix-free description language is a description language. The second inequality follows from the fact that any description d can be turned into a prefix-free description by prefixing it by the following prefix-free description of the length $|d|$ of d: double each bit of the binary representation of $|d|$ and add the string $(0, 1)$ as suffix. □

Remark ([13])**.** We can complement (19) and (20) by $\mathbf{K}'(\sigma) := \min_{L \ni \sigma} K'(L)$, where K' is as defined in Remark 6. We will refer to $\mathbf{K}'(\sigma)$ as the *intermediate time complexity* of σ; notice that

$$\mathbf{C} - O(1) \leq \mathbf{K}' \leq \mathbf{K} + O(1).$$

8 *A priori* time semimeasure

We can also define an analogue of Levin's *a priori* semimeasure (see, e.g., [12], Section 7.33) for time. A *time semimeasure* is a function $P : (\mathbf{X} \times 2)^* \to [0, 1]$ such that, for all infinite data sequences ω,

$$\sum_{l=0}^{\infty} P(\omega^l) \leq 1.$$

Theorem 9. *There is a largest to within a constant factor lower semicomputable time semimeasure.*

Proof. It is easy to check that there exists a sequence P_k, $k = 1, 2, \ldots$, of semicomputable time semimeasures that is *jointly lower semicomputable*, in the sense of the function $(k, \sigma) \mapsto P_k(\sigma)$ being lower semicomputable, and *universal*, in the sense of containing every lower semicomputable time semimeasure. For any such sequence, the average

$$\mathbf{M} := \sum_{k=1}^{\infty} 2^{-k} P_k$$

will be a largest to within a constant factor lower semicomputable time semimeasure. □

Let us fix a largest to within a constant factor lower semicomputable time semimeasure $\mathbf{M}$ and call it the *a priori* time semimeasure. We will use the notation M for the standard *a priori* semimeasure on $\mathbb{N}_0$; it is well known that $-\log M$ coincides with prefix complexity K to within an additive constant (see, e.g., [12], Theorem 7.29). For the time counterparts of M and K we will only state a weaker result.

Theorem 10. $\mathbf{C} - O(1) \le -\log \mathbf{M} \le \mathbf{K} + O(1)$.

Proof. To check the inequality $-\log \mathbf{M} \le \mathbf{K} + O(1)$, it suffices to check that $2^{-\mathbf{K}}$ is a time semimeasure (its lower semicomputability follows from the upper semicomputability of $\mathbf{K}$). Fix an infinite data sequence ω. For each l, let L_l be the simplest, in the sense of $\mathbf{K}$, law of nature containing ω^l. By the definition of a law of nature all L_l are pairwise distinct, and so we have

$$\sum_{l=0}^{\infty} 2^{-\mathbf{K}(\omega^l)} = \sum_{l=0}^{\infty} 2^{-K(L_l)} \le \sum_{L} 2^{-K(L)} \le 1, \tag{24}$$

where the last sum is over all laws of nature L (the last inequality is obvious, but a detailed proof can be found in, e.g., [12], Theorem 7.27).

To check the opposite inequality $\mathbf{C} \le -\log \mathbf{M} + O(1)$, it suffices to define a description language F for laws of nature such that $\min_{L \ni \sigma} C_F(L) \le -\log \mathbf{M}(\sigma) + O(1)$. For each threshold $k \in \mathbb{N}_0$, we can enumerate (in a computable manner) all data sequences σ satisfying $\mathbf{M}(\sigma) > 2^{-k}$ (as $\mathbf{M}$ is lower semicomputable, we will be able to detect $\mathbf{M}(\sigma) > 2^{-k}$ eventually); let $\sigma_1, \sigma_2, \ldots$ be such an enumeration (the sequence $\sigma_1, \sigma_2, \ldots$ can be finite and even empty, as it is for $k = 0$). Order the 2^k binary strings in 2^k lexicographically. For $n = 1, 2, \ldots$: assign to σ_n as its description the smallest element of 2^k that does not serve as description for any of $\sigma_1, \ldots, \sigma_{n-1}$ that is comparable with σ_n w.r. to $\sqsubseteq$ (in particular, σ_1 has $0^k = (0, \ldots, 0)$ as its description). Since, for each infinite data sequence ω, $\mathbf{M}(\omega^l) > 2^{-k}$ holds for at most 2^k (and even $2^k - 1$) ls, we will never run out of descriptions when following this procedure. Define $F(d)$, where $d \in 2^k$, to be the set of all σ having d as their description; by construction, $F(d)$ is a law of nature and F is a description language (remember that the procedure is repeated for all $k \in \mathbb{N}_0$). Since

$$\mathbf{M}(\sigma) > 2^{-k} \Longrightarrow C_F(\sigma) \le k$$

for all $\sigma \in (\mathbf{X} \times 2)^*$ and $k \in \mathbb{N}_0$, we have $C_F \leq -\log \mathbf{M} + 1$ and, therefore, $\mathbf{C} \leq -\log \mathbf{M} + O(1)$. □

In fact, Alexander Shen pointed out that the standard connection between M and K, $K = -\log M + O(1)$, does not carry over to their time counterparts. (Shen's observation is a version of another standard result in the theory of Kolmogorov complexity.)

Theorem 11 (A. Shen)**.** *It is not true that* $\mathbf{K} = -\log \mathbf{M} + O(1)$.

Proof. Suppose that, in fact, $\mathbf{K} = -\log \mathbf{M} + O(1)$. Fix an object $\mathbf{x} \in \mathbf{X}$ and two labels $a, b \in \mathbf{Y}$. Set $A := (\mathbf{x}, a) \in \mathbf{Z}$, $\alpha := (A, A, \ldots) \in \mathbf{Z}^\infty$, and $B := (\mathbf{x}, b) \in \mathbf{Z}$. For each $n \in \mathbb{N}$, consider the following n finite data sequences:

$$\alpha^n B \alpha^1, \quad \alpha^n B \alpha^2, \ldots, \quad \alpha^n B \alpha^n.$$

Since there is a time semimeasure P satisfying $P(\alpha^n B \alpha^k) = 1/n$, for all $n \in \mathbb{N}$ and all $k = 1, \ldots, n$, we have $\mathbf{M}(\alpha^n B \alpha^k) \geq 1/(cn)$, for all $n \in \mathbb{N}$ and all $k = 1, \ldots, n$, c standing for a positive universal constant (with different occurrences of c referring to possibly different positive universal constants). By our assumption, $2^{-\mathbf{K}(\alpha^n B \alpha^k)} \geq 1/(cn)$, for all $n \in \mathbb{N}$ and all $k = 1, \ldots, n$. Remember that $\sum_L 2^{-K(L)} \leq 1$, where the sum is over all laws of nature (we have already used this: see (24)). The series $\sum_L 2^{-K(L)}$ contains at least n terms $2^{-K(L)} \geq 1/(cn)$ (since laws of natures containing $\alpha^n B \alpha^k$ and $\alpha^n B \alpha^{k'}$ are necessarily different when $k \neq k'$). The series is positive, and so its sum will not change if we rearrange its terms. Let us sort them in the decreasing order. The nth largest term will be at least $1/(cn)$, and therefore $\sum_n 1/n = \infty$ implies $\sum_L 2^{-K(L)} = \infty$. This contradiction concludes the proof. □

Remark ([1])**.** As shown by Mikhail Andreev, it is also not true that $\mathbf{K}' = -\log \mathbf{M} + O(1)$, where $\mathbf{K}'$ is intermediate time complexity, as defined in Remark 7. The proof is much more difficult and can be found in [1].

9 Time randomness

In the usual theory of Kolmogorov complexity the notion of algorithmic randomness is as important as that of algorithmic complexity (and perhaps was the main motivation behind Kolmogorov's introduction of algorithmic complexity). There are many versions of algorithmic randomness, and in this paper we will briefly discuss only the time analogue of Kolmogorov's original definition $|\sigma| - C(\sigma)$ of the randomness deficiency of a binary string σ of length $|\sigma|$ (given, somewhat implicitly, in [3], Section 4) and, later on (see Theorem 13), the time analogue of Martin-Löf's [4] definition of randomness.

The *time randomness deficiency* of a finite data sequence $\sigma \in (\mathbf{X} \times 2)^*$ is defined to be

$$\mathbf{D}(\sigma) := \log |\sigma| - \mathbf{C}(\sigma).$$

(We take $\log|\sigma|$ instead of Kolmogorov's $|\sigma|$ in view of Theorem 5: whereas the trivial upper bound on plain Kolmogorov complexity is $C(\sigma) \le |\sigma| + O(1)$, the trivial upper bound on plain time complexity is

$$\mathbf{C}(\sigma) \le C(|\sigma|) + O(1) \le \log|\sigma| + O(1).)$$

Informally, we can rewrite (23) as

$$\mathcal{U}_{\le N} \approx \{\sigma \mid \mathbf{C}(\sigma) \le \log N\}.$$

We could have defined the universal prediction system by

$$\mathcal{U}'_m := \{\sigma \mid \mathbf{C}(\sigma) \le m\}$$

(with m in place of $\log N$). This definition would be especially useful in situations without noise where we can expect to make a finite number of prediction errors over an infinite data sequence. In situations where there is noise at a more or less constant level for each observation (which is typical under the assumption, prevalent in machine learning and nonparametric statistics, that the observations are independent and identically distributed), it may be more useful to replace $\mathbf{C}$ by $\mathbf{D}$ and set, for each threshold $m \in \mathbb{N}_0$,

$$\Delta_m := \{\sigma \mid \mathbf{D}(\sigma) > m\}.$$

The corresponding prediction sets are

$$\Pi_{\Delta_m}(s) := \{y \in 2 \mid (s,y) \notin \Delta_m\} = \{y \in 2 \mid \mathbf{D}((s,y)) \le m\}.$$

In a situation $s = (\sigma, x)$, the prediction system Δ_m predicts that the label $y \in 2$ of x will be an element of $\Pi_{\Delta_m}(s)$. The following simple result shows that the rate at which this prediction system makes errors is less than 2^{-m}.

Theorem 12. *For each infinite data sequence $\omega = ((x_1,y_1),(x_2,y_2),\ldots)$, each $l \in \mathbb{N}$, and each $m \in \mathbb{N}_0$,*

$$\left|\left\{i \in \{1,\ldots,l\} \mid y_i \notin \Pi_{\Delta_m}(\omega^{i-1}, x_i)\right\}\right| = \left|\left\{i \in \{1,\ldots,l\} \mid \omega^i \in \Delta_m\right\}\right| < 2^{-m} l.$$

Proof. If the prediction system Δ_m makes an error when predicting y_i, i.e., $y_i \notin \Pi_{\Delta_m}(\omega^{i-1}, x_i)$, we have $\mathbf{D}(\omega^i) > m$, and so

$$\mathbf{C}(\omega^i) < \log i - m \le \log l - m.$$

The number of such i does not exceed the number of all descriptions of length less than $\log l - m$, i.e., does not exceed $2^{\log l - m} - 1 < 2^{-m} l$. □

In the rest of this section we will explore more systematically prediction systems of the type Δ_m. (Notice that, formally, they are not even weak prediction systems as defined in Section 4.) A *randomness-type prediction system* is a jointly enumerable family Λ of sets $\Lambda_m \subseteq (\mathbf{X} \times 2)^*$ of finite data sequences such that:

- Λ_m are nested: $\Lambda_0 \supseteq \Lambda_1 \supseteq \Lambda_2 \supseteq \cdots$;
- for all $m \in \mathbb{N}_0$, $l \in \mathbb{N}$, and $\omega \in (\mathbf{X} \times 2)^\infty$,

$$\left|\left\{i \in \{1,\dots,l\} \mid \omega^i \in \Lambda_m\right\}\right| \le 2^{-m} l. \tag{25}$$

Theorem 12 says that Δ is a randomness-type prediction system. It is easy to see that there is a universal randomness-type prediction system:

Theorem 13. *There exists a randomness-type prediction system $\mathcal{D}$ such that, for any randomness-type prediction system Λ, there exists $c \in \mathbb{N}$ such that, for all $m \in \mathbb{N}_0$, $\Lambda_{m+c} \subseteq \mathcal{D}_m$.*

Proof. Notice that we can enumerate all randomness-type prediction systems $\Lambda^1, \Lambda^2, \dots$, in the sense that there is a recursively enumerable set

$$\Lambda \subseteq (\mathbf{X} \times 2)^* \times \mathbb{N}^2$$

such that:

1. For any k, the sequence $(\Lambda^k_m)_{m=1}^\infty$, where

   $$\Lambda^k_m := \{\sigma \in (\mathbf{X} \times 2)^* \mid (\sigma, m, k) \in \Lambda\}$$

   is a randomness-type prediction system.
2. Any randomness-type prediction system coincides, for some k, with the sequence $(\Lambda^k_m)_{m=1}^\infty$.

(The existence of such Λ follows from the existence of such a set Λ' when item 1 is ignored and the fact that we can enumerate the elements of Λ' one by one including each of them into Λ if and only if the inclusion does not violate item 1.) We can then combine all these randomness-type prediction systems into $\mathcal{D}$ setting

$$\mathcal{D}_m := \bigcup_{k=1}^\infty \Lambda^k_{m+k}. \tag{26}$$

We will get a randomness-type prediction system, since

$$\begin{aligned}
\left|\left\{i \in \{1,\dots,l\} \mid \omega^i \in \mathcal{D}_m\right\}\right| &= \left|\bigcup_{k=1}^\infty \left\{i \in \{1,\dots,l\} \mid \omega^i \in \Lambda^k_{m+k}\right\}\right| \\
&\le \sum_{k=1}^\infty \left|\left\{i \in \{1,\dots,l\} \mid \omega^i \in \Lambda^k_{m+k}\right\}\right| \\
&\le \sum_{k=1}^\infty 2^{-m-k} l = 2^{-m} l,
\end{aligned}$$

and this system is obviously universal. □

Let us fix a randomness-type prediction system $\mathcal{D}$ satisfying the condition in Theorem 13 and call it the *universal randomness-type prediction system*; set, for any situation s and any $m \in \mathbb{N}_0$,

$$\Pi_{\mathcal{D}_m}(s) := \{y \in 2 \mid (s,y) \notin \mathcal{D}_m\}.$$

A crude connection of $\mathcal{D}$ with our previous definition is given in the following theorem.

Theorem 14. *There exists $c > 0$ such that, for any finite data sequence $\sigma \in (\mathbf{X} \times 2)^{l-1}$ (for any $l \in \mathbb{N}$), any $x \in \mathbf{X}$, and any threshold $m \in \mathbb{N}$,*

$$\Pi^{cl2^{-m}m^2}_{\mathcal{U}}((\sigma,x)) \subseteq \Pi_{\mathcal{D}_m}((\sigma,x)). \tag{27}$$

This theorem asserts that the prediction set output by the universal prediction system is at least as precise as the prediction set output by the universal randomness-type prediction system if we increase slightly the allowed percentage of errors: from 2^{-m} to $c2^{-m}m^2$. It involves not just multiplying by a constant (as in, e.g., (9)) but also the term m^2, which is logarithmic in the allowed percentage of errors 2^{-m} for $\mathcal{D}_m$.

By Theorem 12, Theorem 14 will stay true if we replace the right-hand side $\Pi_{\mathcal{D}_m}((\sigma,x))$ of (27) by $\Pi_{\Delta_m}((\sigma,x))$; moreover,

$$\Pi_{\mathcal{D}_m}((\sigma,x)) \subseteq \Pi_{\Delta_{m+c}}((\sigma,x))$$

for a constant c.

Proof of Theorem 14. Let us replace (27) by the equivalent

$$\sigma \in \mathcal{D}_m \Longrightarrow \sigma \in \mathcal{U}_{\leq c|\sigma|2^{-m}m^2}.$$

Define a prediction system $\mathcal{L} = (L_1, L_2, \ldots)$ as, essentially, $\mathcal{D}_m$; formally:

- The law of nature L_1 contains only finite data sequences $\sigma \in \mathcal{D}_m$ of length at most 2^m. This set if prefix-free by the definition of a randomness-type prediction system: indeed, (25) shows that, for any infinite data sequence ω, at most $2^{-m}2^m = 1$ element of L_1 is a prefix of ω.
- The next 2 laws of nature (L_2 and L_3) contain only finite data sequences $\sigma \in \mathcal{D}_m$ of length in the range $2^m + 1$ to 2^{m+1}, and we will define them similarly to the proof of Theorem 10. Enumerate, in a computable manner, all such data sequences σ ($\sigma \in \mathcal{D}_m$ and $|\sigma| \in [2^m + 1, 2^{m+1}]$); let $\sigma_1, \sigma_2, \ldots$ be such an enumeration. For $n = 1, 2, \ldots$: include σ_n into the law of nature (L_2 or L_3) with the smallest index that does not already contain data sequences comparable with σ_n in the sense of the order $\sqsubseteq$ (in particular, $\sigma_1 \in L_2$). Two laws of nature (L_2 and L_3) are sufficient since, by (25), each infinite data sequence ω has at most $2^{-m}2^{m+1} = 2$ elements of $\mathcal{D}_m$ with length in the range $[2^m + 1, 2^{m+1}]$ (and even $[0, 2^{m+1}]$) as its prefixes.

- The next 4 laws of nature (L_4 to L_7) contain only finite data sequences $\sigma \in \mathcal{D}_m$ of length in the range $2^{m+1}+1$ to 2^{m+2}. Enumerate, in a computable manner, all data sequences $\sigma \in \mathcal{D}_m$ whose length is in this range; let $\sigma_1, \sigma_2, \ldots$ be such an enumeration. For $n = 1, 2, \ldots$: include σ_n into the law of nature (L_4 to L_7) with the smallest index that does not already contain data sequences comparable with σ_n in the sense of the order $\sqsubseteq$. We will never run out of the available laws of nature (L_4 to L_7) by the definition of a randomness-type prediction system: see (25).

- And so on.

Any data sequence $\sigma \in \mathcal{D}_m$ whose length l is in the range $2^{m+i-1}+1$ to 2^{m+i}, $i \in \mathbb{N}$, will be included in one of the 2^i laws of nature L_{2^i} to $L_{2^{i+1}-1}$, and so

$$\sigma \in \mathcal{L}_{\le 2^{i+1}-1} \subseteq \mathcal{L}_{\le 2^{2-m}(l-1)-1}.$$

In combination with Theorem 3, we obtain

$$\sigma \in \mathcal{U}_{\le c_1 2^{K(\mathcal{L})} 2^{-m} l}$$

for a constant $c_1 > 0$. Therefore, our task reduces to checking that

$$2^{K(m)} \le c_2 m^2$$

for a constant $c_2 > 0$. Since $2^{-K(m)}$ is the universal semimeasure on the positive integers (see, e.g., [12], Theorem 7.29), we even have

$$2^{K(m)} \le c_3 m (\log m)(\log\log m)\cdots(\log\cdots\log m),$$

where the product contains all factors that are greater than 1 (see [10], Appendix A). □

Remark. The proof shows that the inclusion (27) can be strengthened to

$$\Pi_{\mathcal{U}}^{cl2^{K(m)-m}}((\sigma, x)) \subseteq \Pi_{\mathcal{D}_m}((\sigma, x)).$$

Next we show how the constant c in Theorem 13 depends on Λ. First we give a standard definition of prefix complexity adapted to randomness-type prediction systems.

A *description language for randomness-type prediction systems* is a function F mapping 2^* to the set of all randomness-type prediction systems such that the set

$$\{(d, \sigma, m) \in 2^* \times (\mathbf{X} \times 2)^* \times \mathbb{N} \mid \sigma \in F_m(d)\}$$

is recursively enumerable, where $F_m(d)$ is the mth set in $F(d)$, i.e., $F_m(d) := \Lambda_m$ when $F(d) = (\Lambda_1, \Lambda_2, \ldots)$. The *effective domain* $\mathrm{dom}(F)$ of a description language F for randomness-type prediction systems is (11). A *prefix-free description language for randomness-type prediction systems* is a description language F for randomness-type prediction systems such that $\mathrm{dom}(F)$ is prefix-free.

The *complexity* of a randomness-type prediction system Λ with respect to a description language F for randomness-type prediction systems is defined by

$$C_F(\Lambda) := \min\{|d| \mid \Lambda = F(d)\}.$$

Analogously to Theorem 2 (but using the fact that we can enforce item 1 on p. 19) we can prove:

Theorem 15. *There is a description language U for randomness-type prediction systems that is universal in the sense that for any description language F for randomness-type prediction systems there exists a constant c such that, for any randomness-type prediction system Λ,*

$$C_U(\Lambda) \le C_F(\Lambda) + c.$$

There is a prefix-free description language U' for randomness-type prediction systems that is universal in the sense that for any prefix-free description language F for randomness-type prediction systems there exists a constant c such that, for any randomness-type prediction system Λ,

$$C_{U'}(\Lambda) \le C_F(\Lambda) + c.$$

We fix a universal description language U for randomness-type prediction systems and call $C(\Lambda) := C_U(\Lambda)$ the *plain complexity* of Λ. And we fix a universal prefix-free description language U' for randomness-type prediction systems and call $K(\Lambda) := C_{U'}(\Lambda)$ the *prefix complexity* of Λ.

Theorem 16. *There exists a constant $c \in \mathbb{N}$ such that, for any randomness-type prediction system Λ and any $m \in \mathbb{N}_0$, $\Lambda_{m+K(\Lambda)+c} \subseteq \mathcal{D}_m$.*

Proof. Let U' be our chosen universal prefix-free description language for randomness-type prediction systems. Analogously to the proof of Theorem 13, we can then combine all randomness-type prediction systems into one system $\mathcal{D}'$ by setting

$$\mathcal{D}'_m := \bigcup_{d \in 2^*} U'_{m+|d|}(d) \tag{28}$$

(cf. (26)). We again get a randomness-type prediction system:

$$\begin{aligned}\left|\left\{i \in \{1,\dots,l\} \mid \omega^i \in \mathcal{D}'_m\right\}\right| &= \left|\bigcup_{d \in 2^*} \left\{i \in \{1,\dots,l\} \mid \omega^i \in U'_{m+|d|}(d)\right\}\right| \\ &\le \sum_{d \in \mathrm{dom}(U')} \left|\left\{i \in \{1,\dots,l\} \mid \omega^i \in U'_{m+|d|}(d)\right\}\right| \\ &\le \sum_{d \in \mathrm{dom}(U')} 2^{-m-|d|} l \le 2^{-m} l.\end{aligned}$$

The inclusion $\Lambda_{m+K(\Lambda)} \subseteq \mathcal{D}'_m$ now follows from $\Lambda = U'(d)$ for some $d \in 2^{K(\Lambda)}$. The addend "$+ c$" allows us to replace the randomness-type prediction system $\mathcal{D}'$ defined by (28) by our chosen universal randomness-type prediction system $\mathcal{D}$. □

In conclusion of this section we will reword our definition of a universal prediction system to make it more similar to that of a universal randomness-type prediction system. A *complexity-type prediction system* is a jointly enumerable family Λ of sets $\Lambda_m \subseteq (\mathbf{X} \times 2)^*$ of finite data sequences such that, for all $m \in \mathbb{N}_0$ and $\omega \in (\mathbf{X} \times 2)^\infty$,

$$\left|\left\{i \in \mathbb{N} \mid \omega^i \in \Lambda_m\right\}\right| \le 2^m. \tag{29}$$

Theorem 17. *There exists a complexity-type prediction system $\mathcal{V}$ such that, for any complexity-type prediction system Λ, there exists $c \in \mathbb{N}$ such that, for all $m \in \mathbb{N}_0$, $\Lambda_m \subseteq \mathcal{V}_{m+c}$.*

Fix a complexity-type prediction system $\mathcal{V}$ satisfying the condition in Theorem 17 and call it the *universal complexity-type prediction system*. The following analogue of Theorem 6 shows that this is not an essentially new notion.

Theorem 18. *There is a constant $c > 0$ such that, for all $m \in \mathbb{N}_0$,*

$$\{\sigma \mid \mathbf{C}(\sigma) \le m - c\} \subseteq \mathcal{V}_m \subseteq \{\sigma \mid \mathbf{C}(\sigma) \le m + c\}. \tag{30}$$

Proof. The left-hand inclusion in (30) is obvious. The right-hand inclusion is witnessed by the following description language for laws of nature. Enumerate in a computable manner all finite data sequences $\sigma_1, \sigma_2, \dots$ in $\mathcal{V}_m$. Order all binary strings in 2^m lexicographically. Assign 0^m (i.e., the sequence $(0, \dots, 0)$ of length m) to σ_1 as its description. For $i = 2, 3, \dots$, assign to σ_i as its description the first string in 2^m that has not being assigned as yet to the strings among $\sigma_1, \dots, \sigma_{i-1}$ that are comparable with σ_i. The finite data sequences with the same description now form a law of nature with that description. Repeat for all $m \in \mathbb{N}_0$.

The only thing that remains to be checked is that we will never run out of strings in 2^m. Let us check this carefully. It is convenient to think of the elements of 2^m as colours, and our goal is to show that we will never run out of the 2^m available colours. We look at the set $(\mathbf{X} \times 2)^*$ of all finite data sequences as a tree (rooted at $\square$ and with σ and σ' connected with an edge when $\sigma \sqsubset \sigma'$ but there is no σ'' such that $\sigma \sqsubset \sigma'' \sqsubset \sigma'$). *Siblings* are non-empty finite data sequences that differ only in their last element. Let us fix some stage i of the construction in the previous paragraph; at the beginning of this stage we have a partial colouring of the tree $(\mathbf{X} \times 2)^*$: the vertices $\sigma_1, \dots, \sigma_{i-1}$ have been coloured, and our task is to colour σ_i. For each vertex σ, let T_σ be the set of all colours used in the tree rooted at σ, and P_σ be the set of all colours used along the path from the root $\square$ to σ (not including σ). Notice the following properties of our construction:

1. The colours of comparable vertices are different.
2. If a vertex σ gets colour d, then each smaller color is used either for a predecessor of σ or for a descendant of σ.
3. If σ is a vertex (coloured or not), then the sets T_σ and P_σ are disjoint (by Property 1), and T_σ is an initial segment in the complement $2^d \setminus P_\sigma$ of P_σ.

Indeed, if a colour appears in T_σ, it is the colour of some vertex $\sigma' \sqsupseteq \sigma$, and so all smaller colours appear either before σ' (therefore, in P_σ or T_σ) or after σ' (therefore, in T_σ).

4. The sets T_σ and $T_{\sigma'}$ for any two siblings σ and σ' are comparable with respect to inclusion. Indeed, they are two initial segments of the same complement.

5. For each vertex σ the total number of colours used in T_σ is minimal, in the sense of being equal to the maximal number of coloured vertices on the paths in T_σ. This can be shown by an inductive argument using the previous property.

The last property, in combination with (29), shows that we will have at least one colour left for σ_i. □

10 Universal conformal prediction under the IID assumption

Up to this point our exposition has been completely probability-free, but in this section we will consider the special case where the data are generated in the IID manner. For basic definitions of the theory of conformal prediction see, e.g., [14]. For simplicity, we will only consider computable conformity measures that take values in the set $\mathbb{Q}$ of rational numbers. Remember that $\mathcal{D}$ is the universal randomness-type prediction system, as introduced in the previous section; let us set $\mathcal{D}_m := (\mathbf{X} \times 2)^*$ for $m < 0$ (i.e., we include all finite data sequences in $\mathcal{D}_m$ for negative m).

Theorem 19. *Let Γ be a conformal predictor based on a computable conformity measure taking values in $\mathbb{Q}$. Then there exists $c \in \mathbb{N}$ such that, for almost all infinite data sequences $\omega = ((x_1, y_1), (x_2, y_2), \ldots) \in (\mathbf{X} \times 2)^\infty$ and all significance levels $\epsilon \in (0, 1)$, from some l on we will have*

$$\Pi_{\mathcal{D}_{\lfloor -\log \epsilon \rfloor - c}}((\omega^{l-1}, x_l)) \subseteq \Gamma^\epsilon((\omega^{l-1}, x_l)). \tag{31}$$

This theorem says that the prediction set output by the universal randomness-type prediction system is at least as precise as the prediction set output by Γ, to within the usual additive constant.

Proof of Theorem 19. Without loss of generality we can and will assume $\epsilon \in (0, 1/2)$. For each such ϵ set $m := \lfloor -\log \epsilon \rfloor - 1$. (Intuitively, we replace ϵ by a new significance level 2^{-m}, which we make at least twice as large as the original ϵ.) Let Λ_m be $\Gamma^{2^{-m}}$ forced to satisfy (25); formally, Λ_m contains only finite data sequences σ such that $\Gamma^{2^{-m}}$ makes an error when predicting the last label in σ, and Λ is defined by induction first on m and then on the length of σ as follows: σ is included in Λ_m if and only if:

- σ is included in all Λ_i, $i < m$ (this condition is satisfied automatically if $m = 1$);
- the condition (25) is satisfied, where l is the length of σ and ω is an infinite continuation of σ.

By the standard validity property of conformal predictors ([14], Corollary 1.1), we will have

$$\Pi_{\Lambda_m}((\omega^{l-1}, x_l)) \subseteq \Gamma^\epsilon((\omega^{l-1}, x_l))$$

from some l on almost surely. □

Remark. The proof shows that we can replace the c in (31) by $c + K(\Gamma)$, where c now does not depend on Γ and $K(\Gamma)$ is the smallest prefix complexity of the programs for computing the conformity measure on which Γ is based.

11 The theory of Kolmogorov complexity

In this section we will discuss the theory of Kolmogorov complexity as a special case of our theory. We obtain the former by taking $\mathbf{X}$ and the label space (2 in this paper) of size one. More generally, the theory of Kolmogorov complexity embeds into our theory when we fix an object and a label and only consider sequences of identical observations with those object and label. Therefore, let us fix an element $\mathbf{x}$ of $\mathbf{X}$ and a label, say 0.

Let $o \in (\mathbf{X} \times 2)^*$ be the infinite data sequence $((\mathbf{x}, 0), (\mathbf{x}, 0), \ldots)$ consisting of identical observations $(\mathbf{x}, 0)$.

Theorem 20.

$$\begin{aligned} \mathbf{C}(o^n) &= C(n) + O(1), \\ \mathbf{K}(o^n) &= K(n) + O(1), \\ -\log \mathbf{M}(o^n) &= -\log M(n) + O(1). \end{aligned} \tag{32}$$

Proof. We will only prove (32); the other two relations can be proved similarly. Reinterpreting a description of $n \in \mathbb{N}$ as a description of the law of nature $(\mathbf{X} \times 2)^n$, we obtain the inequality $\leq$ in (32). (Alternatively, we can notice that (32) is a special case of the inequality $\leq$ of (21).) And reinterpreting a description of a law of nature L as a description of the length of the only element of $L \cap \Sigma(o)$, we obtain the inequality $\geq$ in (32). □

Combining Theorem 20 with the standard fact that $K(n) = -\log M(n) + O(1)$ (e.g., [12], Theorem 7.29), we can see that Theorem 10 can be improved when restricted to $\Sigma(o)$: in this case $-\log \mathbf{M} = \mathbf{K} + O(1)$. Unfortunately, the equality cannot be extended to all finite data sequences: see Theorem 11.

12 Conclusion

In this paper we have ignored the computational resources, first of all, the required computation time and space (memory). Developing versions of our definitions and results taking into account the time of computations is a natural next step. In analogy with the theory of Kolmogorov complexity, we expect that the simplest and most elegant results will be obtained for computational models that are more flexible than Turing machines, such as Kolmogorov–Uspensky algorithms and Schönhage machines.

An interesting open question is whether Theorem 10 can be improved to $-\log \mathbf{M} = \mathbf{K} + O(1)$ by modifying the definition of prefix time complexity (Theorem 11 says that a modification is necessary, and Remark 8 shows that intermediate time complexity does not work). Another open question is whether plain complexity C can be improved to (or almost to) prefix complexity K in Theorem 4.

More open questions are raised by the definition of universal randomness-type prediction systems in Section 9: how can such prediction systems be characterized in terms of other notions (such as plain and prefix time complexity, time randomness deficiency, and *a priori* time semimeasure) introduced in this paper or in terms of similar notions? (In Theorem 14 we gave only the most obvious connection.)

Acknowledgments

We thank the anonymous referees of the conference and journal versions of this paper for helpful comments. In particular, comments made by the referees of the journal version have led to Remarks 2 and 7, and we especially appreciate their generosity in filling a gap in the proof of Theorem 18. This work has been supported by the Air Force Office of Scientific Research (grant "Semantic Completions"), EPSRC (grant EP/K033344/1), and the EU Horizon 2020 Research and Innovation programme (grant 671555).

References

[1] Mikhail Andreev and Alexander Shen. Stopping time complexity, 2017. Unpublished manuscript.

[2] Ronald A. Fisher. *Statistical Methods and Scientific Inference*. Hafner, New York, third edition, 1973.

[3] Andrei N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1:1–7, 1965. Russian original: Три подхода к определению понятия "количество информации".

[4] Per Martin-Löf. The definition of random sequences. *Information and Control*, 9:602–619, 1966.

[5] Cheryl Misak. Charles Sanders Peirce (1839–1914). In Cheryl Misak, editor, *The Cambridge Companion to Peirce*, chapter 1, pages 1–26. Cambridge University Press, Cambridge, 2004.

[6] Charles S. Peirce. The scientific attitude and fallibilism. In Justus Buchler, editor, *Philosophical Writings of Peirce*, chapter 4, pages 285–318. Dover, New York, 1955.

[7] Karl R. Popper. *Logik der Forschung*. Springer, Vienna, 1934. English translation: *The Logic of Scientific Discovery*. Hutchinson, London, 1959.

[8] Karl R. Popper. *Objective Knowledge: An Evolutionary Approach*. Clarendon Press, Oxford, revised edition, 1979. First edition: 1972.

[9] Karl R. Popper. *All Life is Problem Solving*. Abingdon, Routledge, 1999.

[10] Jorma Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11:416–431, 1983.

[11] Glenn Shafer. *The Art of Causal Conjecture*. MIT Press, Cambridge, MA, 1996.

[12] Alexander Shen. Around Kolmogorov complexity: Basic notions and results. In Vladimir Vovk, Harris Papadopoulos, and Alexander Gammerman, editors, *Measures of Complexity: Festschrift for Alexey Chervonenkis*, chapter 7, pages 75–115. Springer, Cham, 2015.

[13] Alexander Shen. Private communications, 2016–2017.

[14] Vladimir Vovk. The basic conformal prediction framework. In Vineeth N. Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk, editors, *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations, and Applications*, chapter 1, pages 3–19. Elsevier, Amsterdam, 2014.

[15] Vladimir Vovk. Continuous-time trading and the emergence of probability. Technical Report `arXiv:0904.4364v4` [math.PR], `arXiv.org` e-Print archive, May 2015. Journal version: *Finance and Stochastics*, 16:561–609, 2012.