Tomasz Steifer

Computable Prediction of Infinite Binary Sequences with Zero-One Loss

Rozprawa doktorska

Promotor: dr hab. Łukasz Dębowski, prof. IPI PAN
Promotor pomocniczy: dr Dariusz Kalociński

Instytut Podstaw Informatyki
Polskiej Akademii Nauk
Warszawa, 2019

Acknowledgements

This dissertation would not exist in its present form without generous support of many people that I had a pleasure of making acquaintance with. To begin with, I thank my mom for raising me in a home full of books which obviously helped me develop a certain curiosity about the world. For *the wonder is the only beginning of philosophy.*

I am forever indebted to the late Marcin Mostowski — the very person who introduced me to the mathematical logic. Marcin had an enormous influence on my scientific development. He was nothing less then a role model — not only of a scientist but also of a person and a loyal friend.

I thank both of my advisors for their help. Dariusz Kalociński not only co-authored several results presented in this report but also supported my aspiration to turn these results into a doctoral dissertation. He also contributed many corrections to this thesis. By the way, any major mistakes in the Chapter II are probably his fault too.

Łukasz Dębowski was kind enough to agree to become my principal advisor. My initial plans for the dissertation did not included the contents of Chapter III. Łukasz virtually made me learn some modern probability theory together with the rudiments of measure theory and ergodic theory. This in turn allowed me to grasp a slightly different perspective on the matter of prediction.

A special thanks to Marcin Lewandowski, my boss at the Institute of Fundamental Technological Research of the Polish Academy of Sciences. I am well aware that Marcin does not share my enthusiasm for the mathematical logic and similar impractical pursuits. Nevertheless, he agreed to treat my research on prediction as a part of my work duties. It would be considerably more difficult to complete this thesis without his support.

I should also thank several friends who gave me encouragement and cheered at my pursuit of academic accolades — Magda, Kasia, Agnieszka, Klaudia, to name but a few. Also many thanks to Marcin and Ziemek for uncounted discussions over a cup(s) of coffee. Thank you all.

Finally, a dedication is due. This dissertation is an attempt in studying the limits of our ability to predict future outcomes of uncertain events. I dedicate it to my great-grandfather, Tomasz Adam Steifer, who according to the family tradition shot himself after loosing considerable amounts of money gambling, and to all those who once made an effort to predict the outcome of an uncertain and failed.

Contents

Streszczenie

Predyktorem nazywamy całkowitą funkcję ze zbioru słów binarnych w zbiór $\{0, 1\}$. Odpowiada to operacji zgadywania $i + 1$-szego bitu na podstawie informacji o pierwszych i zaobserwowanych bitach. Rozprawa skupia się wokół predyktorów, które są jednocześnie całkowitymi obliczalnymi funkcjami.

Miarą skuteczności predyktora jest strata zerojedynkowa, która liczona jest jako stosunek błędnych zgadywań do wszystkich zgadywań dokonanych do pewnego momentu. Wielkość tę nazywamy błędem predykcji. W szczególności, interesuje nas asymptotyczne zachowanie błędów predykcji. Badane są takie własności jak zbieżność, optymalność czy istnienie schematów predykcji, które są uniwersalne dla zadanej klasy procesów. W toku wywodu prezentowane wyniki umiejscowione są w kontekście algorytmicznej teorii losowości.

Abstract

A predictor is a total function from binary words into the set of binary outcomes — $\{0,1\}$. It is interpreted as an operation of guessing $i+1$-th outcome using information about the first i outcomes of some process. This thesis focuses on predictors which are total computable functions.

The performance of a predictor is assessed via the zero-one loss function, that is, we study the ratio between the number of wrong guesses and the total number of guesses made so far. This ratio is called the prediction error. In particular, the thesis deals with questions concerning the asymptotic behaviors of prediction errors. We study such issues as convergence, optimality and existence of schemes that are universal for the class of stationary ergodic processes. The results are presented in the context of the algorithmic randomness theory.

Introduction

Suppose that we deal with a finite but potentially infinite set of data, possibly representing empirical observations about some physical system. A natural question which occurs in this case is the question of prediction. Although prediction might be understood in various different ways, in this thesis we focus on prediction of the nearest future of the underlying system based on its unfolding behaviour. For a very simple example, consider an infinite sequence of binary values. It could be the case that this sequence is a binary representation for some well defined number, e.g. square root of two. If that is so, then this sequence is deterministic and if someone asks us to guess the i-th bit of the sequence, we can always give the right answer, at least in principle.

Suppose, however, that the sequence is a witness to some random phenomena. For example, Rosencrantz is continuously flipping a fair unbiased coin, while Guildenstern writes down the outcomes — 0 for heads, 1 for tails. What can be said about our ability to predict such a sequence? Does the knowledge about past outcomes change our situation in any way? Is there any optimal strategy to approach this problem? Even more, what theoretical framework is the best to describe this scenario? Are we dealing with an individual sequence or rather the whole universe of equipropable sequences? Furthermore, can we go the other way around and infer about the system in question based on its apparent (un)predictability?

To tackle these questions, a mathematical notion of predictors is introduced. A predictor is simply a function that takes a binary word (a prefix of some infinite sequence) as an argument and outputs zero or one. In this thesis, we choose to restrict our attention to binary sequences only. There are various philosophical and practical reasons for this, but these will not be discussed here.

To compare various prediction schemes and to measure their performance, an appropriate loss function is needed. Here, a natural zero-one loss is considered. Formally speaking, at each stage of prediction we study the ratio of incorrect answers to all predictions made so far. It is assumed that we never abstain from making a guess. We will call this measure a prediction error. Since we are dealing with potentially infinite objects, our attention will be focused on the asymptotic behavior of the prediction error. In particular, sequences and predictors for which the error converges in the limit will be of a special interest.

It is important to note, that other loss functions could be considered as well. For example, we could require that the sequence of guesses converges to giving the right answers at some point — in other words, that we are wrong only finitely many times. In the probabilistic setting in turn, various authors use word *prediction* to denote empirical

estimation of uknown conditional probabilities. In such case, one can be interested in different loss functions such as vanishing of quadratic differences between the true probabilities and the estimates. Prediction via predictors and measure estimation are connected in various ways. While it is not the central topic of the present work, some insights on these connections will be provided.

Finally, we impose one important constrain on predictors. Our goal is to describe prediction as it is done (or in principle, as it may be done) in sciences. Therefore, we are not interested in any form of magical prediction familiar to the world of witches, prophets and crystal gazing. We want the prediction to be **effective**. A minimal constrain for effectiveness is given by the mathematical notion of computability. Therefore, we will focus on computable predictors, which will be called *proper predictors*. In the context of theoretical computer science, such idea was — to the best of my knowledge — introduced by Ko [24] (by the name of *total prediction functions*).

Looking at prediction from this perspective, various computability-oriented questions and problems will manifest themselves. In particular, we will ask how predictability (or unpredictability) connects to well studied notions from computability such as Turing degrees.

The structure of this work is as follows. It begins with an introductory part which deals with elementary notions from computability theory and modern probability theory. It is then followed by the second part which deals with the prediction in the deterministic setting — prediction of individual sequences. Three problems are studied in this part.

Firstly, a simple notion of unpredictability, known as Ko stochasticity, is introduced. It is then placed in the context of algorithmic randomness theory. Most of this section follows known results. This part also contributes a novel result by showing that no Ko stochastic sequence may be computably enumerable (in fact, not even n-computably enumerable).

Secondly, we move our attention to sequences for which errors of no predictor converge. These sequences will be called unstable. An unstable c.e. sequence is constructed using the priority method. This is a joint work with Dariusz Kalociński from the paper [23].

Thirdly, we consider the notion of optimality. It is shown that for a certain sequence no predictor is optimal. Again, this theorem comes from a joint paper with Dariusz Kalociński [23]. This part is ended by some simple observations regarding the connections between predictors and predictability as defined by Tadaki [36] as well as connections with coarse computability, which is one of the most fruitful recent developments in computability.

Eventually, we move on to the probabilistic framework in the third part of this dissertation. Following probabilistic results on universal prediction for stationary ergodic processes, we provide effective versions of several theorems. In particular, it is shown that there exists such a proper predictor that is optimal on every sequence random in the Martin-Löf sense relative to some computable stationary ergodic measure. In this way we fill up another gap between the standard probabilistic framework and its algorithmic counterpart. A probabilistic version of the optimality theorem from the second

part is also provided. This theorem is a refined version of the result presented in Paris on the IEEE International Symposium on Information Theory 2019 [22].

Notation

By convention, the term *sequences* refers to infinite objects constructed over the alphabet $\{0,1\}$. The set of finite words over $\{0,1\}$ is denoted by $2^{<\mathbb{N}}$. The set of all one-sided infinite sequences is denoted by $2^{\mathbb{N}}$. The set of all two-sided infinite sequences is denoted by $2^{\mathbb{Z}}$. For $k > 0$, the set of words of length k is sometimes denoted by 2^k.

Following information-theoretic convention, bits of a one-sided infinite sequence are indexed from 1. It is assumed that $0 \in \mathbb{N}$. The set of all nonzero natural numbers is denoted as $\mathbb{N}^+$ (similarly, $\mathbb{R}^+$ denotes the set of positive real numbers). The empty word is denoted by $\square$.

Given a sequence (or a word) x we denote i-th bit of x by x_i. To denote a string $x_j, x_{j+1}, \ldots, x_k$ (with $j < k$) we write x_j^k. In particular, a prefix of length n of x is denoted by x_1^n. By convention, for a sequence (or a word) x, we let x_1^0 denote the empty word as well.

$|\sigma|$ is used to denote the length of the word σ, while $\#X$ refers to the cardinality of set X. $|x|$ may also denote the absolute value if $x \in \mathbb{R}$.
$\#_1$ is used to denote the number of nonzero bits in a word, that is

$$\#_1\sigma = \#\{i < |\sigma| : \sigma_i = 1\}.$$

We sometimes write $\sigma(\prec) \preceq \tau$ to say that σ is a (proper) prefix of τ.

We identify sets of natural numbers with binary sequences in the natural way: every n-th bit is nonzero if and only if n is in the corresponding set. Similarly, we sometimes assume correspondence between words and sets: $\sigma_n = 1$ if and only if n is in the corresponding set and no $k > |\sigma|$ belongs to the set. In this manner we sometimes apply set-theoretic notation to sequences in words. In particular, $\sigma \subseteq \tau$ means that if $\sigma_i = 1$ then $\tau_i = 1$, while $\omega = \bigcup_{i=1}^{\infty} \sigma_i$ (with $\sigma_i \in 2^{<\mathbb{N}}$) means that $\omega_i = 1$ if and only if for some i we have $\sigma_i = 1$.

Given sets A and B, the set $C = A \oplus B$ is constructed by putting bits of A on odd bits of C and bits of B on even bits C.

For a function f, $dom(f)$ denotes the domain of f, while given set A we write $f[A]$ to denote the image of A. Following a standard set-theoretic notation, we let $\mathcal{P}(\Omega)$ denote the power set of a set Ω.

Part I

Preliminaries

1

Computability theory

1.1 Computable functions and sets

Partial computable functions may be defined using various equivalent models of computation [11]. For completeness, a certain definition will be also provided here. This definition will be based on register machines.

The register machines consist of the unbounded number of registers which are unrestricted in their capacity and a set of admissible instructions. A program for the register machine defines which registers will be used (a finite number of these) and which instructions are to be executed and in what order. A register is functionally equivalent to a variable in modern programming languages, with the exception of its unlimited capacity as it can store an arbitrary large natural number.

Definition 1 (Register machine)**.** *The register machine consists of registers* $r_1, r_2, \ldots$ *A program for the register machine consists of a finite number of instructions coming from the set of admissible instructions. For each* $i, j, k \in \mathbb{N}$ *we have the following instructions:*

- *zeroing of* j*-th register* $(r_j := 0)$ *— the value stored in the register* r_j *is changed to zero; all other registers remain unchanged;*
- *succesor for* j*-th register* $(r_j := r_j + 1)$ *— the value stored in the register* r_j *is increased by* 1*; all other registers remain unchanged;*
- *copying of* j*-th register to* i*-th register* $(r_i := r_j)$ *— the value stored in the register* r_i *is replaced by the value stored in the register* r_j*; all other registers (including* r_j*) remain unchanged;*
- *conditional jump* $(IF\ r_j = r_i\ GO\ TO\ k)$ *— if the values stored in* j*-th and* i*-th register are equal, execute the* k*-th instruction. Otherwise, execute the next instruction.*

Since a program is finite list of instructions, each program uses only a finite number of registers. Suppose that the register machine is executing some program and i *is the*

maximal index of a register used in this program (i.e., no register r_j with $j > i$ is used). We encode the configuration of the machine on program P after some number of steps as $(s_0, s_1, \ldots, s_i)$ where s_0 is the index of the instruction that should be executed in the next step and $s_1, \ldots, s_i$ are equal to values stored by registers $r_1, \ldots, r_i$ respectively, with $r_1, \ldots, r_i$ being all the registers used by the program P.

We say that a $(v_1, v_2, \ldots)$ is a computation of the register machine on program P if for every j the states v_j and v_{j+1} are related according to the instruction executed in the $(j+1)$-th step and $v_1 = (0, s_1, \ldots, s_k, 0, \ldots, 0)$ for some k. We will say that $s_1, \ldots, s_k$ is an input of the register machine on program P.

Let b be the total number of instructions in the program P. We will say that the machine stops on program P with input $s_1, \ldots, s_k$ if there exists a computation with a configuration $v_m = (y_0, y_1, \ldots, y_k)$ such that $y_0 > b$. Moreover, we will say that y_1 is the output of the computation.

Definition 2 (Partial computable function)**.** *We say that $f : \mathbb{N}^c \to \mathbb{N}$ is a partial computable function if there exists a program P such that for every $x_1, \ldots, x_c \in \mathbb{N}$ there exists $y \in \mathbb{N}$ such that*

$$f(x_1, \ldots, x_c) = y$$

if and only if the computation of P with input $x_1, \ldots, x_c$, stops. Moreover, if this computation stops, then the output is y. We will write $f(x_1, \ldots, x_c) \downarrow$ if $f(x_1, \ldots, x_c)$ is defined and $f(x_1, \ldots, x_c) \uparrow$ otherwise.

We say that f is computable if it is total, that is $f(n)$ is defined for every natural n.

Observation 3. *There are countably many partial computable functions.*

Proof. Observe that every partial computable function corresponds to a program and every program is a finite object and may be represented by a natural number. □

When talking about functions from a countable set to a countable set, we often identify objects from countable sets with natural numbers. In particular, binary words may be identified with natural numbers using the standard binary representation. Similarly, a positive rational number $q \in \mathbb{Q}^+$ may be identified with its representation $(a, b) \in \mathbb{N}^2$

$$q = \frac{a}{b}.$$

In this manner, the definition of partial computable functions (and other definitions introduced later) may be generalized for functions from a countable set to a countable set — via an appropriate coding.

In particular, an arithmetical decision problem of form:

$$\text{given } x \text{ is } \phi(x) \text{ satisfied?}$$

where ϕ is arithmetical formula, may be seen as equivalent to a set

$$\{x \in \mathbb{N} : \phi(x)\}.$$

Definition 4 (Characteristic function)**.** *Let $A \subseteq \mathbb{N}^k$. We will say that $\chi_A : \mathbb{N}^k \to \{0, 1\}$ is a characteristic function of A if and only if for every $\bar{x} \in \mathbb{N}^k$ we have $\chi_A(\bar{x}) = 1$ if and only if $\bar{x} \in A$.*

Definition 5 (Computable sets)**.** *A set is computable*[1] *if its characteristic function is computable.*

This gives us an important notion of computability of sets or problems. A problem X is computable if there is an algorithm which allows us to decide, in a finite time, for each natural number whether it belongs to X or not. An early realization of Alan Turing was that not every problem is computable. Before defining a canonical example of an uncomputable problem, we introduce the notion of a universal program. We start by noting that all programs may be enumerated in an effective way (since they are finite objects which may be effectively coded — for details see e.g. Chapter 4 of [11]). We fix $h_1, h_2, \ldots$ — a canonical enumeration of partial computable functions with only one argument. The halting set is defined as

$$\{x \in \mathbb{N} : h_x(x) \downarrow\}.$$

Theorem 6 (Turing [39])**.** *The halting set is not computable.*

The proof is based on a simple diagonal argument which is possible by the following observation. There exists a program which can simulate an arbitrary program if given the code of the program as input. Indeed, we can prove that

Theorem 7. *Let $h_1, h_2, \ldots$ be a canonical enumeration of partial computable functions with only one argument. There exists a program U such that for all $i \in \mathbb{N}$, $h_i(x)$ is defined if and only if $U(i, x, 0, \ldots, 0) \downarrow$ and if $h_i(x) = y$, then the program U ends its computation on input i, x with the output y.*

Proof. [11], Chapter 4, gives the proof of analogous result for Turing machines. One can easily adapt this reasoning for the purpose of the register machine. □

We will call such program U universal. We know that there is no effective procedure that will decide for an arbitrary natural number, whether it belongs to the halting set or not. That being said, there is a procedure by which every element of the halting set (and no other) is listed after some finite time, as the following reasoning asserts.

Observe that we can simulate all programs simultaneously in the following effective way. Fix an effective enumeration of programs. At each n-th step we simulate n first steps of the computation of the first n programs with their codes as inputs. We have only a countable set of programs and a countable set of steps. A product of countable sets is countable. Hence, if some program halts then it will halt sooner or later in our simulation, and consequently we can list effectively all the elements of the halting set.

This observation lead us to the notion of computable enumerability.

Definition 8 (Computably enumerable sets)**.** *A set A is computably enumerable (c.e.) if there is a partial computable function f such that $dom(f) = A$.*

[1]For historical reasons, 'computable' is sometimes used interchangeably with 'recursive'.

Proposition 9. *If A is c.e. and $A \neq \emptyset$, then there exists a total computable function which enumerates A, i.e., A is the image of the function.*

Proof. If A is c.e. then there exists a partial computable function f such that $dom(f) = A$ and the corresponding program P for the register machine. Now, the enumeration is straighforward. Simply simulate P on all possible inputs in an effective way. Each time that $f(n)\downarrow$, output n as a part of the enumeration. □

Finally, recall that we have a natural correspondence between sets and binary sequences, namely, given a set X there is a sequence Y for all $n > 0$, n if and only if $Y(n) = 1$. This simple observation allow us to talk about computability of sequences, c.e. sequences etc.

1.2 Turing reductions

Computability gives us a certain demarcation line between problems. However, it is rather crude. For various reasons, we might want to have a more fine grained classification of problems. For computable problems, this may be done by considering upper bounds for computation time or the number of registers used (time and space complexities). But this idea does not work for uncomputable problems, whereas these will be of interest to us. Turing's idea was to consider an ordering imposed on problems by the relation of computable reducibility. This has a very natural interpretation. Roughly speaking, A is reducible to B, simply if knowing the answer to B allows us to get the answer to A. A formal definition follows.

Definition 10 (Oracle machine)**.** *A program for the register oracle machine consists of a finite number of instructions coming from the set of admissible instructions. The set of admissible instructions is contain all instructions from Definition* 1 *joined with the following instruction:*

- *oracle inquiry (IF $r_i \in X$ GO TO k) — if the value stored in i-th register is a member of set X, execute the k-th instruction. Otherwise, proceed to the instruction as ordered in the program.*

Note that the oracle inquiry is a syntactical object — it's just an another instruction. Depending on what set is actually supplied as the oracle (i.e., what set is being denoted by X), the answer to $r_i \in X$ might be different.

All auxiliary notions (computation, stoping etc.) are defined in an analogous way as in the case of a regular register machine. In particular, we will have the notion of a set being computable with an oracle. By convention, we say X is Y-computable if X is computable with the oracle Y.

Definition 11. *We say that A is Turing reducible to B (or that A is computable in B) if A is computable with the oracle B. We also write $A \leq_T B$ to say that A is Turing reducible to B. We say that problems A and B are Turing equivalent $A \equiv_T B$ if $A \leq_T B$ and $B \leq_T A$.*

Observation 12. *The relation $\equiv_T$ is an equivalence relation.*

Definition 13 (Turing degree)**.** *A Turing degree is an equivalence class of the relation $\equiv_T$.*

Definition 14 (Turing jump)**.** *A Turing jump $\mathbf{a}'$ of a degree $\mathbf{a}$ is a set of indices i in the canonical enumeration $h_1^X, h_2^X, \ldots$ of programs for the register machines with an oracle $X \in \mathbf{a}$, such that $h_i^X(i, 0, \ldots) \downarrow$.*

In particular, we mention the degree $\mathbf{0}$ which corresponds to computable problems. The halting set is then in the degree $\mathbf{0}'$.

1.3 Arithmetical hierarchy

Furthermore, we introduce the arithmetical hierarchy which classifies sets of natural numbers via their corresponding first order definitions.

Definition 15 (Arithmetical hierarchy)**.** *We define classes Σ_n^0 and Π_n^0 inductively for relations on natural numbers.*

1. *Firstly, let both Σ_0^0 and Π_0^0 denote the class of computable relations.*
2. *For $n \geq 0$, relation ψ belongs to Σ_{n+1}^0 if and only if there exists a relation ϕ in class Π_n^0 such that*

$$\psi(y_1, \ldots, y_j) \Leftrightarrow \exists x_1, \ldots, x_k \phi(y_1, \ldots, y_j, x_1, \ldots, x_k).$$

3. *For $n \geq 0$, relation ψ belongs to Π_{n+1}^0 if and only if there exists a relation ϕ in class Σ_n^0 such that*

$$\psi(y_1, \ldots, y_j) \Leftrightarrow \forall x_1, \ldots, x_k \phi(y_1, \ldots, y_j, x_1, \ldots, x_k).$$

We will say that relation is Σ_n^0 or Π_n^0 meaning that it belongs to class Σ_n^0 or Π_n^0 respectively. Similarly, we will say that a set $A \subset \mathbb{N}$ is Σ_n^0 or Π_n^0 if it is defined by a relation ψ of class Σ_n^0 or Π_n^0 respectively, that is

$$\forall x \in \mathbb{N} : x \in A \Leftrightarrow \psi(x).$$

We will say that a relation or set is Δ_n^0 if it is both Σ_n^0 and Π_n^0.

Clearly, if a set A is computable then it is Σ_1^0.

Theorem 16 (Arithmetical hierarchy)**.** *A set A is computably enumerable if and only if it is Σ_1^0.*

Sketch. Suppose that A is Σ_1^0. That is for each y

$$y \in A \Leftrightarrow \exists x_1, \ldots, \exists x_k \phi(y, x_1, \ldots, x_k)$$

where k is natural and ϕ is a computable predicate. Since k is finite, we can effectively enumerate every possible combination of $k+1$ natural numbers. The algorithm for enumeration of y works as follows. Enumerate all tuples of length $k+1$. For each tuple $z_0, \ldots, z_k$ check if $\phi(z_0, \ldots, z_k)$. If so, z_0 is an element of A. This gives an effective enumeration of A which shows that A is computably enumerable.

The proof for the other direction is a slightly more involved. For simplicity, we sketch the proof for unary relations only. One need to define the so-called Kleene predicate T (invented by Stephen C. Kleene). Roughly speaking, $T(a, b, c)$ states that c is a code for the computation history of the program with index a starting with input b and terminating with a halting state. Since every halting computation is a finite iteration of finite number of operations, one can code it as natural number. Moreover, assuming a fixed coding procedure one can effectively check if a given number is a code of a halting computation. Suppose that given input x, the a-th program enumerates A and halts if and only if it finds out that $x \in A$. Such program exists if A is computably enumerable. Now, we can give a simple Σ_1^0 definition for A:

$$\forall x \in \mathbb{N} : x \in A \Leftrightarrow \exists c T(a, x, c).$$

□

Observation 17. *If a set A is Σ_n^0 for some n, then its complement is Π_n^0.*

Observation 18. *If a set A is Σ_n^0 or Π_n^0 for some n, then it is Σ_m^0 and Π_m^0 for every $m > n$.*

1.4 Computability and real functions

A significant part of mathematics deals with real numbers. Some of real numbers (i.e., their digits) are easy to describe by an algorithm. On the other hand, we already know that there are only countably many algorithms, while the set of real numbers is uncountable. Hence, some real numbers will be beyond our algorithmic comprehension. To formalize this intuition, we introduce computability notions into the realm of real numbers.

Definition 19 (Computable reals)**.** *We say that a $r \in \mathbb{R}$ is computable if and only if the left cut of r, that is, $\{q \in \mathbb{Q} : q < r\}$ is computable.*

In other words, the real is computable if it is approximable by some computable sequence of rationals such that we can control the approximation error. We get a weaker notion if the latter condition is lifted.

Definition 20 (Left-c.e. reals)**.** *We say that a $r \in \mathbb{R}$ is left computably enumerable (left-c.e.) if and only if the left cut of r is computably enumerable.*

Definition 21 (Uniform computability)**.** *A sequence of reals is uniformly computable if and only if the corresponding sequence of left cuts is uniformly computable, that is, if there exists a computable function* $g : \mathbb{N} \times \mathbb{Q} \to \{0, 1\}$ *such that for the* i*-th real* r *in the sequence and for every* q*,* $g(i, q) = 1$ *if and only if* q *belongs to the left cut of* r*.*

Definition 22 (Uniform computable enumerability)**.** *A sequence of reals is uniformly c.e. if and only if the corresponding sequence of left cuts is uniformly c.e., that is, if there exists a computable function* $g : \mathbb{N} \times \mathbb{N} \to \mathbb{Q}$ *such that for the* i*-th real* r *in the sequence,* $g(i, j)$ *enumerates the left cut of* r*.*

Definition 23 (Computable real-valued functions)**.** *A function* f *from a countable domain (such as* $\mathbb{N}$ *or* $2^{<\mathbb{N}}$*) to* $\mathbb{R}$ *is computable if and only if its values are uniformly computable.*

Definition 24 (Left-c.e. real-valued functions)**.** *A function* f *from a countable domain (such as* $\mathbb{N}$ *or* $2^{<\mathbb{N}}$*) to* $\mathbb{R}$ *is left-c.e. if and only if its values are uniformly left-c.e.*

2

Probability theory

2.1 Measures and probability

For completeness, we recall the fundamental notions of modern probability theory. For a textbook introduction, see e.g., [8].

Definition 25 (π-system, algebra)**.** *Let Ω be a set. A collection of subsets $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ is called a π-system if $\Omega \in \mathcal{F}$ and $\mathcal{F}$ is closed under finite intersection. A π-system is called an algebra if it closed under complement.*

Definition 26 (σ-algebra)**.** *An algebra $\mathcal{F}$ is called σ-algebra if it is closed under countable union i.e., for any subsets $A_i \in \mathcal{F}$ where $i \in \mathbb{N}$, we have*

$$\bigcup_{i \in \mathbb{N}} A_i \in \mathcal{F}.$$

A pair $(\Omega, \mathcal{F})$, where $\mathcal{F}$ is a σ-algebra on Ω, will be called a measurable space.

Definition 27 (generated σ-algebra)**.** *We say that a σ-algebra $\mathcal{F}$ is generated by the class $\mathcal{A} \subseteq \mathcal{P}(\Omega)$, written in symbols as $\mathcal{F} = \sigma(\mathcal{A})$, if $\mathcal{F}$ is the intersection of all σ-algebras $\mathcal{J}$ such that $\mathcal{A} \subseteq \mathcal{J}$.*

It may also useful to consider the following notion

Definition 28 (λ-system)**.** *Let Ω be a set. A collection of subsets $\mathcal{F} \subset \mathcal{P}(\Omega)$ is called an λ-system if $\Omega \in \mathcal{F}$ and $\mathcal{F}$ is closed under complement and countable union of pairwise disjoint sets.*

λ-systems and π-systems are connected via fundamental result by E. Dynkin

Theorem 29 (Dynkin's $\pi - \lambda$ theorem)**.** *Let $\mathcal{A}$ be a π-system and $\mathcal{D}$ be λ-system. If $A \subseteq D$ then the σ-field generated by $\mathcal{A}$ is contained in $\mathcal{D}$.*

Proof. See e.g., Theorem 3.2. in [8]. □

This theorem will be used later on, when we will be interested in probabilities conditioned on infinite past. Before we get there, we will start by defining the standard probability measures.

Definition 30 (finite measure)**.** *Let $\mathcal{F}$ be σ-algebra. Let function $\rho : \mathcal{F} \to \mathbb{R}^{\geq 0}$ be called a finite measure on $\mathcal{F}$ if for any subsets $A_i \in \mathcal{F}$ where $i \in \mathbb{N}$ and $A_i \cap A_j = \emptyset$ for $i \neq j$, we have*

$$\rho\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} \rho(A_i).$$

Definition 31 (probability measure)**.** *A finite measure ρ on the σ-algebra $\mathcal{F}$ of Ω is called a probability measure if $\rho(\Omega) = 1$.*

Definition 32 (probability space)**.** *A tuple $(\Omega, \mathcal{F}, \rho)$ is called a probability space if ρ is a probability measure on σ-algebra $\mathcal{F}$ of Ω.*

It turns out that the σ-algebra equal to the power set $\mathcal{P}(\Omega)$ is usually too large to support interesting probability measures. This happens in particular for $\Omega = 2^{\mathbb{N}}$, being the set of all infinite binary sequences. In this case, we usually consider some specific σ-algebra $\mathcal{F} = \mathcal{2}^{\mathbb{N}}$, called the Borel σ-algebra. Its construction is as follows. For a word $w \in 2^{<\mathbb{N}}$, we will use the following notation to denote the set of its infinite extensions, called a cylinder set:

$$[\![w]\!] := \{x \in 2^{\mathbb{N}} : x_1^{|w|} = w\}.$$

Similarly, for a set of words $S \subset 2^{<\mathbb{N}}$ we introduce the notation:

$$[\![S]\!] = \bigcup_{w\in S} [\![w]\!].$$

The Borel σ-algebra is the σ-algebra generated by the class of all cylinder sets:

$$\mathcal{2}^{\mathbb{N}} := \sigma(\{[\![w]\!] : w \in 2^{<\mathbb{N}}\}).$$

Probability measures on σ-algebra $\mathcal{2}^{\mathbb{N}}$ can be defined via the notion of a premeasure:

Definition 33 (premeasure)**.** *Let function $p : 2^{<\mathbb{N}} \to \mathbb{R}^{\geq 0}$ be called a premeasure if $p(\Box) = 1$ and for every $w \in 2^{<\mathbb{N}}$ we have $p(w) = p(w0) + p(w1)$.*

We can move from premeasures to probability measures and back.

Proposition 34 (Kolmogorov process theorem I)**.** *For any premeasure p on $2^{<\mathbb{N}}$ there exists a unique probability measure P on $\mathcal{2}^{\mathbb{N}}$ such that for all words $w \in 2^{<\mathbb{N}}$ we have*

$$P([\![w]\!]) = p(w). \tag{2.1}$$

Conversely, for any probability measure P on $\mathcal{2}^{\mathbb{N}}$ there exists a unique premeasure p on $2^{<\mathbb{N}}$ such that (2.1) holds for all $w \in 2^{<\mathbb{N}}$.

Proof. See Chapter 7 of [8]. □

We will say that probability measure P is generated by premeasure p if condition (2.1) holds. An important example of a measure is the uniform measure. The uniform measure λ is generated by premeasure $u(w) = 2^{-|w|}$. This measure may be naturally interpreted as representing probabilities of independent unbiased coin tossing. The unbiasedness condition means that for each toss the probability of the outcome being a tail is 1/2. The independence condition means that the conditional probability of the next outcome does not depend on the outcome of previous one. It is easy to arrive at a conclusion that the probability of the first k bits being equal to some word w is equal to 2^{-k}. One can also note that the uniform measure corresponds to the Lebesgue measure Λ on interval $[0, 1]$. If for a sequence $x \in 2^{\mathbb{N}}$ we define $\phi(x) := \sum_{i=1}^{\infty} x_i 2^{-i}$ then for a word $w \in 2^{<\mathbb{N}}$ we obtain that $\phi([\![w]\!])$ is an interval of length $2^{-|w|}$. Consequently, for any Borel set $A \in \mathcal{2}^{\mathbb{N}}$ we have $\Lambda(\phi(A)) = \lambda(A)$. For this reason, the uniform measure λ is sometimes called the Lebesgue measure.

For practical reasons we will take often abuse the notation and identify premeasures with their corresponding measures. In particular, if μ is a measure with some corresponding premeasure ρ, we will write $\mu(w)$ to denote $\rho(w)$ (or equivalently — $\mu([\![w]\!])$) and so on.

By convention, if $\mu(A) = 1$, we will say that A happens μ-almost surely (or simply, almost surely). Similarly, if

$$\mu(\{\omega : \phi(\omega)\}) = 1,$$

we will say that $\phi(\omega)$ for almost every ω.

2.2 Computable measures

We will find the notion of a premeasure quite useful when trying to incorporate computability-theoretic notions into the world of probabilities. One can observe that some measures are computationally easy (e.g., unbiased coin tosses), while other somehow escape our limited comprehension (e.g., biased coin tosses with an uncomputable probability of a tail). We want to explicate that intuition via the notion of a computable measure. A standard definition makes use of the fact that every premeasure is a real function. Hence:

Definition 35 (computable measure)**.** *We will say that a measure μ is computable if the corresponding premeasure is a computable real function.*

Prediction may be seen as guessing which bit is more probable as the outcome of some future experiment. The following proposition tells us that computability does not guarantee that we can compute the answer to the question which bit is more probable (if any).

Proposition 36. *There is a computable measure μ such that the set*

$$\{\sigma : \mu(\sigma 0) = \mu(\sigma 1)\}$$

is not computable.

Proof. Fix some enumeration of machines $M_1, M_2, \ldots$ and consider f defined for each $n \in \mathbb{N}$ by

$$f(n) = 0.c_1c_2\ldots$$

where $c_i = 1$ if and only if the machine M_n halts on input n after i steps. The above definition is stated in decimal notation for natural numbers. Note that $f(n) = 0$ if an only if M_n never halts on n.

We define a measure μ as follows. For each $\sigma \in 2^{<\mathbb{N}}$ let

$$\mu(\sigma 0) = \mu(\sigma)(1/2 + f(|\sigma|)).$$

We can well approximate $\mu(\sigma 0)$ simply by simulating M_n. But if the set $\{\sigma : \mu(\sigma 0) = \mu(\sigma 1)\}$ is computable, then the halting problem is computable as well, which constitutes a contradiction.

□

2.3 Random variables

In this section we will recall the basic concept of a random variable and its expectation.

Definition 37 (measurable function)**.** *Let $(\Omega, \mathcal{F})$ and $(\mathbb{X}, \mathcal{X})$ be measurable spaces. A function $f : \Omega \to \mathbb{X}$ is called a measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{X}, \mathcal{X})$ if $A \in \mathcal{X}$ implies*

$$f^{-1}[A] \in \mathcal{F}.$$

A measurable function $T : \Omega \to \Omega$ is called a measurable transformation.

Definition 38 (random variable)**.** *Let $(\mathbb{X}, \mathcal{X})$ be a measurable space. A random variable X with an image space $(\mathbb{X}, \mathcal{X})$ on a measure space $(\Omega, \mathcal{F}, \mu)$ is an arbitrary measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{X}, \mathcal{X})$.*

As we proceed our studies, we will become interested in assigning probabilities to propositions, that is, given a predicate ϕ we would like to know the measure of events that satisfy ϕ. Thus, we adopt a notational convention which reflects the natural correspondence between propositions and random variables with the set of logical values {true, false} as their image:

$$(\phi) := \{\omega \in \Omega : \phi(\omega)\},$$

$$\mu(\phi) := \mu(\{\omega \in \Omega : \phi(\omega)\}).$$

By convention, we use $\mathbf{1}$ to denote the indicator variable i.e., given a predicate ϕ

$$\mathbf{1}\{\phi(\omega)\} = \begin{cases} 1 & \text{if } \phi(\omega) \quad (2.2) \\ 0 & \text{otherwise.} \quad (2.3) \end{cases}$$

Definition 39 (discrete and simple random variables)**.** *A random variable X with an image space $(\mathbb{X}, \mathcal{X})$ is called a discrete random variable if the set $\mathbb{X}$ is countable and $\mathcal{X}$ is generated by $\{\{x\} : x \in \mathbb{X}\}$. A discrete random variable is called simple if its image $\mathbb{X}$ is finite.*

Definition 40 (stochastic process)**.** *A stochastic process X is a collection of random variables X_i with $i \in I$ which share the same image space $(\mathbb{X}, \mathcal{X})$. We say that the process is one-sided infinite (or simply—infinite) if $I = \mathbb{N}^+$ and two-sided infinite if $I = \mathbb{Z}$.*

In this dissertation, our attention will orbit around binary processes as defined below.

Definition 41 (binary random variables and processes)**.** *A random variable is called binary if its image is $\mathbb{X} = \{0, 1\}$. A process consisting of binary random variables will be called a binary process.*

In our applications, a typical stochastic process $X = X_1, X_2, \ldots$ $(X = \ldots X_{-1}, X_0, X_1, \ldots)$ consists of projection functions mapping a one-sided infinite (two-sided infinite) binary sequence ω to its i-th bit ω_i, i.e., $X_i(\omega) = \omega_i$. In the case of these projections, if X is a process on the measure space $(\Omega, \mathcal{F}, \mu)$, we will say that μ is a probability distribution of X. Random variables and stochastic processes provide us with yet another notation for cylinder sets generated by words $w \in 2^{<\mathbb{N}}$. For every $k \in \mathbb{Z}$ and $n = |w|$ let

$$(X_{k+1}^{k+n} = w) := (X_{k+1} = w_1 \wedge X_{k+2} = w_2 \wedge \ldots \wedge X_{k+n} = w_n).$$

In particular,

$$(X_1^n = w) = [\![w]\!].$$

Definition 42 (real variable)**.** *A random variable X is called real if its image space is equal to $(\mathbb{R} \cup \{-\infty, \infty\}, \mathcal{R}_\infty)$ where $\mathcal{R}_\infty$ denotes the σ-algebra generated by the set of all semiclosed (possibly infinite) intervals, that is the set*

$$\{[a, b] : a, b \in \mathbb{R} \cup \{-\infty, \infty\}\}.$$

Definition 43 (Lebesgue integral)**.** *For a real random variable Y on a finite measure space $(\Omega, \mathcal{F}, \mu)$ we define the Lebesgue integral $\int Y d\mu$ as*

1. *If $Y \geq 0$ and Y is a simple random variable:*

$$\int Y d\mu = \sum_{y : \mu(Y=y)>0} y\mu(Y = y).$$

2. *If $Y \geq 0$ and Y is not a simple random variable:*

$$\int Y d\mu = \sup_{X \leq Y} \int X d\mu,$$

where the supremum is taken over all real simple random variables X satisfying $X \leq Y$.

3. If $Y = Y_+ - Y_-$, where $Y_+, Y_- \geq 0$ and either $\int Y_+ d\mu < \infty$ or $\int Y_- d\mu < \infty$, then:

$$\int Y d\mu = \int Y_+ d\mu - \int Y_- d\mu.$$

Definition 44 (expectation)**.** *Assuming a fixed probability measure μ, we define the expectation $\mathbb{E}_\mu$ of the random variable X as*

$$\mathbb{E}_\mu(Y) = \int X d\mu.$$

Typically, when probability measure μ in question is clear from context, we omit the index and write simply $\mathbb{E}(Y) := \mathbb{E}_\mu(Y)$. An elementary property of the expectation follows from the properties of Lebesque integral:

Lemma 45 (Jensen's inequality)**.** *Let $f : (a, b) \to \mathbb{R}$ be a convex measurable function and Y be a real random variable. If $\mathbb{E}(f(Y))$ and $\mathbb{E}(Y)$ are defined then*

$$\mathbb{E}(f(Y)) \geq f(\mathbb{E}(Y)).$$

2.4 Conditional probability

In the elementary probability calculus, the conditional probability of an event A given an event B (with $\mu(B) > 0$) is defined as

$$\mu(A|B) = \mu(A \cap B)/\mu(B).$$

This definition may be roughly interpreted as an answer to the following question—what probability should we assign to A, if we know that B is true? Suppose that we perform an experiment to observe whether B holds or not. This experiment induces a partition γ of the event space:

$$\gamma = \{B, \bar{B}\}.$$

Let us rephrase the previous question—how will the probability assigned to A change depending on the outcome of the experiment? Going further, we might want to think about the conditional probability as a random variable—the conditional probability of A defined with respect to partition γ as

$$\mu(A|\gamma)(\omega) = \mathbf{1}\{\omega \in B\}\mu(A|B) + \mathbf{1}\{\omega \in \bar{B}\}\mu(A|\bar{B})$$

Naturally, more complicated partitions (experiments) may be considered. This dissertation concerns prediction of infinite binary processes. In our framework, prediction is conditioned on the past outcomes of some process. Indeed, most of the time we will try to decide what is the next bit assuming that the previous bits are such and such. Hence, we will often deal with conditional probabilities such as

$$\mu(X_n = 1|X_j^{n-1} = w)$$

for some $w \in 2^{<\mathbb{N}}$ and $j < n$. Such conditional probability gives an answer to the question—how the outcome of experiments $X_j, X_{j+1}, \ldots, X_{n-1}$ influences the probability of X_n? Here, the conditional probability is defined with respect to an appropriate collection of cylinder sets. Ultimately, we will become interested in assigning probabilities relative to some infinite past. Obviously, for a nonatomic measure the set of extensions of an infinite past $x_{-\infty}^{-n}$ (with $x \in 2^{\mathbb{Z}}$), i.e., event

$$(X_{-\infty}^{-n} = x_{-\infty}^{-n})$$

is of measure zero. Hence, the elementary definition cannot be applied here. But as we are about to see, the conditional probability may be generalized to such cases. Indeed, we will try to generalize the notion of conditional probability to conditional probabilities with respect to an arbitrary σ-algebra.

To motivate the general definition let us focus further on the special case. Suppose that we have a binary process X with the probability distribution μ and assume that the values of $X_1, \ldots, X_n$ are to be observed. For simplicity, assume that $\mu(X_1^n = w) > 0$ for all $w \in 2^{<\mathbb{N}}$ of length n. We want to have a notion of probability of an event A (such as $X_{n+1} = 1$) conditioned on the random variable X_1^n. Let us denote it by $\mu(A|X_1^n)$. This entity will be a random variable itself. A minimal constraint for such a notion is following—we expect

$$\mu(A|X_1^n) = \mu(A|X_1^n = w)$$

to hold every time that $X_1^n = w$, where $\mu(A|X_1^n = w)$ is given by the elementary definition

$$\mu(A|X_1^n = w) = \mu(A \wedge X_1^n = w)/\mu(X_1^n = w).$$

To give it a more concrete form, we let

$$\mu(A|X_1^n)(\omega) := \sum_{w \in 2^n} \mathbf{1}\{\omega_1^n = w\}\mu(A|X_1^n = w).$$

In particular, this random variable will satisfy

$$\mu(A \wedge X_1^n = w) = \int_{[\![w]\!]} \mu(A|X_1^n) d\mu$$

for all $w \in 2^n$. Moreover, it will be measurable with respect to the σ-algebra generated by cylinders $[\![w]\!]$, where $|w| = n$.

We will treat these two properties as a more abstract characterization of elementary conditional probability with respect to events of positive measure. To extend this to a general setting, we need to introduce the following measure-theoretic result

Theorem 46 (Radon-Nikodym theorem)**.** *Let μ and ν be finite measures on a measurable space $(\Omega, \mathcal{G})$ such that for every $G \in \mathcal{G}$*

$$\mu(G) = 0 \Rightarrow \nu(G) = 0.$$

Then there exists a function $d\nu/d\mu : \Omega \to [0, \infty)$ satisfying:

i) $d\nu/d\mu$ *is* $\mathcal{G}$*-measurable,*

ii) For every $G \in \mathcal{G}$*,*

$$\nu(G) = \int_G \frac{d\nu}{d\mu} d\mu$$

where $\int_G f d\mu = \int \mathbf{1}\{\omega \in G\} f d\mu$.

Furthermore, any two functions satisfying these conditions are equal μ*-almost surely.*

Proof. See Section 6 of [8]. □

Function $d\nu/d\mu$ is called the Radon-Nikodym derivate.

Now, consider $\mu_A(B) := \mu(A \cap B)$ for an event A. Note that for every event B,

$$\mu(B) = 0 \Rightarrow \mu_A(B) = 0.$$

Hence, the following definition is justifiable.

Definition 47 (conditional probability)**.** *Consider a probability space* $(\Omega, \mathcal{J}, \mu)$ *and a* σ*-algebra* $\mathcal{F} \subset \mathcal{J}$*. For a set* $A \in \mathcal{J}$ *we define conditional probability of* A *relative to* $\mathcal{F}$ *as*

$$\mu(A|\mathcal{F}) = \frac{d\mu_A|_{\mathcal{F}}}{d\mu|_{\mathcal{F}}},$$

where $d\mu|_{\mathcal{F}}$ *is a restriction of* μ *to subdomain* $\mathcal{F}$*.*

Note that the conditional probability is, in fact, a random variable. By the Radon-Nikodym theorem, every two functions satisfying Definition 47 are equal on some set of measure one. That being said, there might be many such functions. Every such function will be called a version of conditional probability. These versions may disagree only on events of zero probability. Hence, in many situations it does not matter which version we have in mind. We note in passing, that in some circumstances (i.e. statistical inference) we may need versions to agree on individual points. This happens under certain topological conditions, for details see [38].

In our work, we will be interested in probabilities conditioned on some fixed past. Possible information about the past observation, e.g., information about X_1^n, will correspond to the σ-algebra $\mathcal{F}$ generated by the collection of cylinders $(X_1^n = w)$ for some w. Then, by convention for an A

$$\mu(A|X_1^n) := \mu(A|\mathcal{F})$$

and given $a, b, c \in \mathbb{Z}$ and an arbitrary $\omega \in 2^{\mathbb{Z}}$ such that $\omega_a^b = w$

$$\mu(X_c = 1|X_a^b = w) := \mu(X_c = 1|X_a^b)(\omega).$$

Note that this value is a constant. For completeness, we note that this conditional probability, defined with respect to σ-algebra generated by a set of cylinders, reduces to the elementary ratio on events of positive measure. Indeed, it may be proven that

Theorem 48. *Let $\mathcal{P}$ be a π-system and let $\mathcal{F}$ be a σ-algebra on Ω generated by $\mathcal{P}$. Suppose that Ω is a finite or countable union of sets from $\mathcal{P}$. An integrable function f is a version of $\mu(A|\mathcal{F})$ if it is measurable with respect to $\mathcal{F}$ and*

$$\int_G f d\mu = \mu(A \cap G)$$

holds for every $G \in \mathcal{P}$.

Proof. See, e.g., Theorem 10.4 in [8]. The Theorem follows directly from it. □

Similarly to the conditional probability, we can define the conditional expectation of some event Y relative to a σ-algebra $\mathcal{F}$.

Definition 49 (conditional expectation)**.** *Consider a probability space $(\Omega, \mathcal{J}, \mu)$ and a σ-algebra $\mathcal{F} \subset \mathcal{J}$. For a real random variable Y (with $Y \geq 0$ and $\mathbb{E}(Y) < \infty$) we define conditional expectation of Y relative to $\mathcal{F}$ as*

$$\mathbb{E}(Y|\mathcal{F}) = \frac{d\mu_Y|_{\mathcal{F}}}{d\mu|_{\mathcal{F}}}$$

where $\mu_Y(A) = \int_A Y d\mu$.

Again, the conditional expectation is a random variable. Among many interesting properties of the conditional expectation one may mention in particular the following:

Lemma 50 (conditional Jensen's inequality)**.** *Let $f : (a, b) \to \mathbb{R}$ be a convex measurable function and Y be a real random variable. If $\mathbb{E}(f(Y)|\mathcal{G})$ and $\mathbb{E}(Y|\mathcal{G})$ are defined then*

$$\mathbb{E}(f(Y)|\mathcal{G}) \geq f(\mathbb{E}(Y|\mathcal{G}))$$

almost surely.

We end this section with an essential comment on the (un)computability of conditional probabilities. Ackerman et al. [1] constructed a pair (X, Y) of computable random variables such that the conditional probability $\mu(X|Y)$ encodes the halting problem (hence, is not computable). However, this does not pose a problem in our circumstances. The prediction algorithms described further in the text will deal with probabilities conditioned on a finite past. As this reduces to the elementary definition on events of interest (i.e., on events of positive measure), they are computable almost everywhere as long as the underlying probability measure is computable. Indeed, the ratio of two computable functions is computable.

2.5 Martingales

As we have already signalled, the abstract measure-theoretic definition of conditional probability was introduced to allow us to talk about probabilities conditioned on events of zero measure (e.g., on the infinite past). However, any example of conditional

probability introduced in the previous section simply reduces to the elementary notion. In this section, we will argue that the abstract definition indeed provides us with a reasonable notion of probability conditioned on the events of zero probability. We are about to see that such measure is, in fact, a limit of probabilities conditioned on events of positive measure. The intuitive idea is that we start with a very coarse partition of the event space and then gradually move through finer and finer partitions. The desired conditional probability corresponds to the fixed point of that procedure. To show that this makes any formal sense, we need to introduce the notion of a martingale process.

Definition 51 (martingale process)**.** *Let $X = X_1, X_2, \ldots$ be a stochastic process on $(\Omega, \mathcal{J}, \mu)$ and let $\mathcal{F}_1, \mathcal{F}_2, \ldots$ be a sequence of σ-algebras in $2^{\mathbb{N}}$ or $2^{\mathbb{Z}}$. X is called a martingale process relative to the σ-algebras $\mathcal{F}_1, \mathcal{F}_2, \ldots$ if the following conditions hold:*

i) $\mathcal{F}_i \subset \mathcal{F}_{i+1}$ *for all* $i \in \mathbb{N}^+$*;*

ii) X_n *is measurable in* $\mathcal{F}_n$ *for all* $n \in \mathbb{N}^+$*;*

iii) $\mathbb{E}(|X_n|) < \infty$ *for all* $n \in \mathbb{N}^+$*;*

iv) $\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n$ *for all* $n \in \mathbb{N}^+$ *almost surely.*

The sequence $\mathcal{F}_1, \mathcal{F}_2, \ldots$ satisfying i) is called a filtration.

The general intuition behind the notion of a martingale process is as follows. Suppose that X_n represents the amount of capital available to the gambler after n-th play. This capital is always finite by the third condition. The filtration $\mathcal{F}_1, \mathcal{F}_2, \ldots$ represents the increasing information—in our special case, the information about outcomes of the first several plays. The fourth condition is simply a fairness condition—the expected value of capital after the next play is equal to the present capital.

On the other hand, we want to think that some filtrations correspond to a progression of finer and finer partitions of the event space, as described at the beginning of this section. It follows from the fundamental result of the martingale theory, the Doob's convergence theorem, that the limit of probabilities conditioned on the elements of this filtration always exists.

Theorem 52 (Doob's convergence theorem)**.** *Let X be a martingale proces relative to the σ-algebras $\mathcal{F}_1, \mathcal{F}_2, \ldots$. Then almost surely the limit*

$$\lim_{n\to\infty} X_n$$

exists and is finite.

Proof. See Theorem 35.5. in [8]. □

A corollary of Doob's convergence theorem tells us that

Theorem 53 (Lévy's law)**.** *Let $\mathcal{F}_1, \mathcal{F}_2, \ldots$ be a filtration and define $\mathcal{F}_\infty$ to be σ-algebra generated by the union $\bigcup_{n\in\mathbb{N}^+} \mathcal{F}_n$. Let $\mathbb{E}(|Y|) < \infty$. Then*

$$\mathbb{E}(Y|\mathcal{F}_\infty) = \lim_{n\to\infty} \mathbb{E}(Y|\mathcal{F}_n)$$

almost surely.

Proof. Consider a filtration $\mathcal{F}_1, \mathcal{F}_2, \ldots$ and the σ-algebra $\mathcal{F}_\infty$ to be σ-algebra generated by the union $\bigcup_{n\in\mathbb{N}^+} \mathcal{F}_n$. Consider a random variable Y and observe that for each n

$$\mathbb{E}(\mathbb{E}(Y|\mathcal{F}_{n+1})|\mathcal{F}_n) = \mathbb{E}(Y|\mathcal{F}_n)$$

almost surely, wheres $\mathbb{E}(|\mathbb{E}(Y|\mathcal{F}_n)|) \leq \mathbb{E}(|Y|) < \infty$, by the conditional Jensen's inequality. Therefore, a collection of random variables $\mathbb{E}(Y|\mathcal{F}_n)$ is a martingale process relative to $\mathcal{F}_1, \mathcal{F}_2, \ldots$. Hence, Doob's convergence theorem may be applied. It may be concluded that the limit

$$\lim_{n\to\infty} \mathbb{E}(Y|\mathcal{F}_n)$$

exists and is finite.

It remains to be shown that

$$\lim_{n\to\infty} \mathbb{E}(Y|\mathcal{F}_n) = \mathbb{E}(Y|\mathcal{F}_\infty).$$

By the definition of conditional expectation and by Radon-Nikodym theorem, it suffices to show that for every $A \in \mathcal{F}_\infty$ we have

$$\int_A \mathbb{E}(Y|\mathcal{F}_\infty)d\mu = \lim_{n\to\infty} \int_A \mathbb{E}(Y|\mathcal{F}_n)d\mu. \tag{2.4}$$

By the dominated convergence theorem, the equation 2.4 holds for any $i \in \mathbb{N}$ and any set $A \in \mathcal{F}_i$. Consequently, it holds for every set $A \in \bigcup_{n\in\mathbb{N}^+} \mathcal{F}_n$. In general, this union is not necessarily a σ-algebra. That being said, we can observe that it is a π-system. Furthermore, the collection $\mathcal{F}$ of sets $A \in \mathcal{F}_\infty$ satisfying equation 2.4 forms a λ-system. Indeed, the fact that $\mathcal{F}$ is closed under countable disjoint union again follows from the dominated convergence theorem. By Dynkin's $\pi - \lambda$ theorem, $\mathcal{F}_\infty$ is contained in $\mathcal{F}$. This ends the proof.

□

Consequently, we can define the probability conditioned on an infinite past as a limit of probabilities conditioned on finite past:

$$\mu(X_0 = 1|X_{-\infty}^{-1} = x_{-\infty}^{-1}) := \lim_{n\to\infty} \mu(X_0 = 1|X_{-n}^{-1} = x_{-n}^{-1}).$$

This value is a concrete version of the conditional probability given the infinite past.

2.6 Useful theorems

Finally, we introduce some basic theorems, which are used as standard tools in probability theory. Unless stated otherwise, the proofs for these may be found e.g., in [8].

Theorem 54 (Markov's inequality)**.** *Let X be a nonnegative real random variable. Then for all $t > 0$*

$$\mu(X \geq t) \leq \frac{1}{t}\mathbb{E}(X).$$

Proof. Fix a nonnegative random variable X with values in $\mathbb{R}$ and $t > 0$. Consider a random variable $Y = X/t$. Then

$$\mu(X \geq t) = \mu(Y \geq 1) = \int_{Y \geq 1} d\mu \leq \int_{Y \geq 1} Y d\mu \leq \int Y d\mu = \frac{1}{t}\mathbb{E}(X).$$

□

Theorem 55 (Hoeffding's inequality [19])**.** *Let $X_1, \ldots, X_n$ be independent binary random variables with the probability distribution μ. Then*

$$\mu(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) > t) \leq e^{-2nt^2}.$$

A more general version of Hoeffding's lemma may be stated for a martingale process.

Theorem 56 (Azuma's inequality [5])**.** *Let $X = X_1, X_2, \ldots$ be a martingale process with values in $\mathbb{R}$, relative to the filtration $\mathcal{F}_1, \mathcal{F}_2, \ldots$ and suppose that $|X_{i+1} - X_i| \leq c$ for all i. Then, for all real $s > 0$ and $i > 1$*

$$\mu\left(|X_{i+1} - X_1| > s\right) \leq 2e^{-s^2/2ic^2}.$$

Recall the interpretation of X_n as representing a gambler's capital after nth play. The Azuma's inequality states that with high probability this capital is close to the starting capital X_0. Moreover, a large divergence from the initial capital will become less probable as the game progresses in time.

The remaining propositions are well-known.

Theorem 57 (monotone convergence theorem)**.** *Let $Y = Y_1, Y_2, \ldots$ be a sequence of nonnegative and nondecreasing real random variables. Then*

$$\sup_{n\in\mathbb{N}} \int Y_n d\mu = \int \sup_{n\in\mathbb{N}} Y_n d\mu.$$

Theorem 58 (dominated convergence theorem)**.** *Let $(Y_n)_{n\in\mathbb{N}}$ be a sequence of real random variables such that $|Y_n| < Z$, where $\int Z d\mu < \infty$. If $\lim_{n\to\infty} Y_n$ exists then*

$$\lim_{n\to\infty} \int Y_n d\mu = \int \lim_{n\to\infty} Y_n d\mu.$$

Theorem 59 (Borel-Cantelli lemma)**.** *Let $A_1, A_2, \ldots$ be an infinite collection of events such that*

$$\sum_{n=1}^{\infty} \mu(A_n) < \infty.$$

Then μ-almost surely only finitely many of A_i happens, i.e. $\mu(\bigcap_{n\in\mathbb{N}} \bigcup_{i\geq n} A_i) = 0$

3

Predictors

Recall that we want to see the prediction as a process, presumably an effective one, which takes a finite number of past observations and makes a guess based on these. In what follows, we give a formal definition of predictor, followed by a formal definition of the prediction error.

Definition 60 (predictor)**.** *A predictor is a total function $f : 2^{<\mathbb{N}} \to \{0, 1\}$. We say that f is a proper predictor if it is a total computable function. Otherwise, we say that it is an improper predictor.*

Definition 61 (prediction error)**.** *Let f be an (im)proper predictor and let $\sigma \in 2^{<\mathbb{N}}$ be a non-empty word, $|\sigma| > 0$. The prediction error of f on σ is defined as*

$$\varsigma(f, \sigma) := \frac{\#\{1 \leq i \leq |\sigma| : \sigma_i \neq f(\sigma_1^{i-1})\}}{|\sigma|} \tag{3.1}$$

For an infinite sequence $x \in 2^{\mathbb{N}}$, the prediction errors of f on A are defined as

$$\varsigma_+(f, x) := \limsup_{n\to\infty} \varsigma(f, x_1^n), \tag{3.2}$$

$$\varsigma_-(f, x) := \liminf_{n\to\infty} \varsigma(f, x_1^n), \tag{3.3}$$

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

We may start with a very simple observation:

Proposition 62. *Fix $A \in 2^{\mathbb{N}}$ and let f be a predictor for A. Suppose that $\varsigma(f, A) = r$. Then, there exists a predictor g for A such that $\varsigma(g, A) = 1 - r$.*

Proof. Given A and f, let g be such that for every τ we have $g(\tau) = 0$ if and only if $g(\tau) = 1$. □

The existence of a relatively good predictor entails the existence of a corresponding, relatively bad predictor (a kind of an *evil twin*).

Part II

Prediction of individual sequences

4

Unpredictability and stochasticity

As if you didn't know how it feels to lose at dice with fate. (...)
As if it wasn't a lifetime spent on connecting the dots,
there was no pattern.

– Mgła, *Exercises in Futility VI*

4.1 Unpredictability via predictors

A natural class of unpredictable sequences may be defined in the framework of predictors. It was first introduced by Ko [24] and later studied by Ambos-Spies et al. [3].

Definition 63. *We say that $x \in 2^{\mathbb{N}}$ is Ko stochastic iff for every proper predictor f we have $\varsigma(f, x) = \frac{1}{2}$.*

Ko stochastic sequences are naturally related to a known notion of stochasticity, namely Church stochasticity. We start by recalling its definition.

Definition 64. *A selection rule is a partial function $f : 2^{<\mathbb{N}} \to \{yes, no\}$. Given a sequence x and a selection rule f, we denote the n-th number k such that $f(x_1^k)$=yes as $s_f(x, n)$. Moreover, if $s_f(x, n-1)$ is defined, let $S_f(x, n) = \sum_{i<n} x_{s_f(x,i)}$ denote the number of ones in the first $n-1$ bits selected by f from x.*
For every f we say that the subsequence selected by f is balanced if and only if

$$\lim_{n\to\infty} \frac{S_f(x, n)}{n} = \frac{1}{2}.$$

Definition 65. *Let $\mathcal{F}$ be a collection of selection functions. A sequence x is stochastic with respect to $\mathcal{F}$ if and only if for every selection rule f from $\mathcal{F}$ either $\lim_{n\to\infty} \frac{S_f(x,n)}{n} = \frac{1}{2}$ (i.e. the subsequence selected by f is balanced) or the number of bits selected by f is finite. We say that a sequence is Church stochastic if it is stochastic with respect to all computable functions.*

Before a construction of a Church stochastic sequence is presented, let as reflect on the relation between Church and Ko stochasticity.

Proposition 66. *Every Church stochastic sequence is Ko stochastic.*

Proof. Let f be a proper predictor. Consider two selection rules s_1 and s_2. The first one selects an i-th bit if and only if f predicts that i-th bit is zero. Otherwise, the second one selects that bit. Now, suppose that x is Church stochastic. By the definition, the subsequences selected by s_1 and s_2 are both infinite and balanced or one is infinite and balanced and one is finite. It follows that $\varsigma(f, x) = \frac{1}{2}$. □

On the other hand, not every Ko stochastic sequence is Church stochastic. Indeed, consider a Ko stochastic sequence with bits of $\emptyset$ (i.e. all-zero sequence) placed on every 2^i-th bit (in fact, every sufficiently fast growing function will be good). Such sequence is not Church stochastic (we have a selection rule that gives an unbalanced sequence $\emptyset$) but the unbalanced subsequence is placed sparsely enough so that it does not influence the error of any predictor. For completeness, a direct construction of such sequence is presented below. There are several ways to construct a Ko stochastic sequence—here, we use the version of the proof of Ville's theorem as presented in [40]. Ville's example provides a method for building a sequence which is stochastic with respect to any countable collection of selection rules. It then remains to recall that every predictor corresponds to two selection rules. Below, we show a generalized version which proves a kind of Church stochasticity relative to some subsequence selected by a function g (only those bits selected by g are used to check if some f selects a balanced subsequence). This will allow us to construct sequences with an unpredictable subsequence (that is, where for any predictor the ratio of correct and wrong prediction on the subsequence approaches $\frac{1}{2}$). The proof is a simple modification of the construction of Ville's sequence, see: section 6.2.2 in [40].

Theorem 67 (Ville [42])**.** *For every countable collection $\mathcal{F}$ of selection functions, there exists a sequence A such that for every f from $\mathcal{F}$ the subsequence selected by f is balanced.*

This construction is usually presented in two steps. Firstly, a version for finite collection of rules is explained and then it is extended to infinite version. Suppose we have a finite collection of selection functions $f_0, f_1, \ldots, f_{k-1}$. We want to construct a sequence x. It will be a mixture of a finite number of subsequences of form $010101\ldots$.

At each step we will be dealing with a subcollection of selection functions. Suppose that x_1^n is already defined. We will deal with selection functions f such that $f(x_1^n) = \text{yes}$. These are active selection functions and each combination of active selection functions corresponds to a binary word σ of length k in the following way: $\sigma_j = 1$ if and only if f_j selects $n+1$-th bit of the sequence. When a combination of selection functions is active and these are coded by σ, we will simply say that σ is active.

There are 2^k words that may be active. As a matter of fact, 2^k is also the number of subsequences $010101\ldots$ which the resulting sequence x will be composed of. We fix a one-to-one correspondence between copies of $010101\ldots$ and words. This correspondence

is used in the construction: the places on which a given subsequence $010101\ldots$ is put depend, among other things, on whether the correponding word σ is active.

Again, suppose that x_1^n is already defined and let σ be a word that codes the active subcollection of selection functions. Let l be a number of times that σ was active before in our construction. Then, we let $x_{n+1} = y_{l+1}$ where $y = 010101\ldots$. Hence, whether we put 0 or 1 depends solely on whether we have dealt with the active combination of functions odd or even number of times.

Suppose that x_1^n is already defined and each combination of selection functions was active a nonnegative number of times. If a combination coded by σ was active an even number of times, then it caused us to put the same number of zeros and ones into x. Otherwise, the difference between these is at most 1. Consequently, at each step the difference between the number of ones and the number of zeros in a prefix x_1^n is bounded by the number of possible combinations of selection functions i.e., by 2^k.

Similarly, for a subsequence selected by a function f, the difference between the number of ones and the number of zeros in a prefix is bounded by the number of all possible combinations of selection functions which include f i.e., by 2^{k-1}. Since k is fixed, this difference become negligible for sufficiently large n and so, every f selects a balanced subsequence of x.

Obviously, this is not enough for the infinite case. We will start with a finite subcollection of selection rules and then gradually add more rules. The trick is to do it slowly enough, so that the number of *active* selection rules is small enough compared to the length of the *active* prefix.

Proof. Let $\mathcal{F} = f_0, f_1, \ldots$ be a countable collection of selection rules. We will construct a sequence x satisfying the condition from the theorem. Let $x_1 = 0$. We will fix a sequence of natural numbers $n_0, n_1, \ldots$ which will be used to provide some lower bounds for lengths of prefixes. We will require these numbers to grow fast e.g., $n_i = 2^{2^i}$. Suppose that for some n we have already defined x_1^n. Let $y \in 2^{\mathbb{N}}$ be such that for every i

$$y_i = 1 \text{ iff } f_i(x_1^n) = \text{yes}.$$

Now, the active word will be the least prefix y_1^m such that y_1^m was active less than n_m times. Similarly as in finite case, each word corresponds to a different copy of $010101\ldots$

Given x_1^n suppose that σ is active on x_1^n. We will let $x_{n+1} = 1$ if and only if the exact number of times σ was active before is odd.
Now, consider a sequence z being a subsequence of x selected by a selection function f_j. Each bit of z corresponds to some prefix of an infinite sequence y which codes selection functions that select this bit.

Suppose that k is the length of the longest σ that contributed to a bit of z_1^n. Without a loss of generality, we may assume that $k > j$ (as we are interested in the asymptotic behavior). Therefore, the number of different words contributing to some bit of x_1^n is bounded from above by $1 + 2 + \ldots + 2^k \leq 2^{k+1}$. This means that at most 2^{k+1} copies of $010101\ldots$ were used to build the prefix z_1^n. Therefore, we know that the number of ones in z_1^n is bounded from below by $\frac{n}{2} - 2^{k+1}$. It is also not possible that the number of ones is greater than the number of zeros (since 1 is always preceded by at least one

0).
On the other hand, we know that a word of length k may be active only if its prefix was active n_{k-1} times. Note that every time a word becomes active a new bit of z is constructed. Thus, we now that $n > n_{k-1}$. And therefore, we get an upper bound for the deviation of the ratio of zeros and ones from 1/2, that is

$$|\frac{1}{2} - \frac{2^{k+1}}{n_{k-1}}|.$$

It remains to see that if the sequence n_i grows fast enough, then x will be balanced for every function from $\mathcal{F}$. □

Corollary 68. *Fix $x \in 2^{\mathbb{N}}$. For every countable collection $\mathcal{F}$ of selection functions, there exists a sequence y such that for every f from $\mathcal{F}$ the subsequence selected by f is balanced. Moreover, for every $i \in \mathbb{N}$ such that y_i is not selected by any function from $\mathcal{F}$, we have $y_i = x_i$.*

Proof. This may be proved by simple modification of the construction of Ville's sequence. The sequence y is constructed by induction. Every time we deal with a bit selected by some function from $\mathcal{F}$ we proceed as with the regular construction for Ville's sequence. If for some $i \in \mathbb{N}$ we have a bit y_i that is not selected by any admissible function we let $y_i = x_i$. □

Corollary 69. *There exists a Ko stochastic sequence which is not Church stochastic*

Proof. We proceed as in Ville's construction with one restriction. Every time we want to define an i-th bit of a sequence x, if $i = 2^j$ for some j the Ville's procedure is bypassed and let $x_i = 0$. Observe that it does not influence prediction error of any predictor (since indexes of form 2^j are sparse), so that the resulting sequence is Ko stochastic. On the other, consider a selection function f such that $f(x_1^n) =$ yes if and only if $n = 2^j$ for some j. The subsequence selected by f is not balanced. Therefore, x is not Church stochastic. □

4.1.1 Martin-Löf tests and randomness

In the previous section we have introduced the class of Church stochastic sequences. Church proposed his notion of stochasticity as an effectivization of von Mises notion of 'collectives' [41]. Following Ville, we have provided a direct construction of such a sequence x. However, we have done it in such a way that for every $k \in \mathbb{N}$,

$$\#_0(x_1^k) \geq \#_1(x_1^k).$$

This is a bit disappointing, if we hoped that the definition would somehow correctly explicate the intuitive notions of randomness or stochasticity: we have a sequence that passes some tests (checking the normality of effective subsequences) but also fails some. One could try to construct a Church stochastic sequences x with a more 'random' ratio between 0s and 1s—for example, satisfying the law of iterated logarithm. We could

then require that all stochastic sequences must satisfy both normality and the law of iterated logarithm. But then again, maybe some other apparent property of stochastic sequences would not be satisfied by x? This would mean that we need again extend our definition and this could go on forever. A natural remedy for such a situation would be to look for a kind of a fixed point.

To do accomplish that, Martin-Löf defined a notion of an effective statistical test for randomness and required that a random sequence passes all such test. It is based on two assumptions. Firstly, we require that a random sequence should not have any rare property. This intuition is formalized by the notion of being of measure zero (with respect to the uniform measure λ). Secondly, the tests of rare properties are to be effective and the effectivization is understood in terms of computability theory.

Definition 70 (uniformly Σ_1^0 collection of subsets of $2^{\mathbb{N}}$)**.** *A collection $U_0, U_1, \ldots$ of sets of sequences is uniformly Σ_1^0 if and only if there is a uniformly c.e. collection $V_0, V_1, \ldots \subset 2^{<\mathbb{N}}$ such that $U_i = [\![V_i]\!]$ for every $i \in \mathbb{N}$.*

Definition 71 (Martin-Löf test)**.** *An uniformly Σ_1^0 sequence $U_0, U_1, \ldots$ of subsets of $2^{\mathbb{N}}$ is called a Martin-Löf test if there exists a computable f such that $\lim_{n\to\infty} f(n) = 0$ and $\lambda(U_n) \leq f(n)$ for every $n \in \mathbb{N}$.*

Definition 72 (Martin-Löf randomness)**.** *A sequence $x \in 2^{\mathbb{N}}$ is called 1-random (or Martin-Löf random) if there is no such Martin-Löf test $U_0, U_1 \ldots$ that $x \in \bigcap_{i\in\mathbb{N}} U_n$.*

This definition may be relativized by replacing computable enumerability with enumerability with respect to some oracles. Thus, a hierarchy of notions of randomness is acquired. This explains how the name 1-randomness comes in place—Martin-Löf randomness forms the first level of an infinite hierarchy of notions of randomness.
As tests are based on computability, one can observe that there is a test that is universal in a sense that it covers every other Martin-Löf test, and suffices to prove or disprove 1-randomness.

Definition 73 (universal Martin-Löf test)**.** *Let $U = U_0, U_1, \ldots$ be a Martin-Löf test. We will say that U is a universal Martin-Löf test if the following holds for each sequence x:*

$$x \notin \bigcap_{i\in\mathbb{N}} U_i$$

if and only if x is 1-random.

Proposition 74 (Martin-Löf [26])**.** *There exists a universal Martin-Löf test.*

Proof. See e.g. Theorem 6.2.5 in [12]. □

One can note an easy but important corollary

Corollary 75. *λ-almost every sequence is 1-random.*

To prove various properties of 1-randomness, it is often advantageous to consider the notion of a Solovay test.

Definition 76 (Solovay test)**.** *A Solovay test is a collection $S_0, S_1, \ldots$ of uniformly Σ^0_1 sets of sequences such that $\sum_{n\in\mathbb{N}} \lambda(S_n) < \infty$. A sequence $x \in 2^{\mathbb{N}}$ is called Solovay random if for every Solovay test, x belongs only to finitely many S_n.*

Theorem 77 (Solovay, unpublished)**.** *A sequence is 1-random if and only if it is Solovay random.*

Proof. See e.g., Theorem 6.2.8 in [12]. □

4.1.2 1-randomness and stochasticity

A natural observation can be made about the relation between Ko stochasticity and 1-randomness, namely, that every 1-random sequence is Ko stochastic. This can be shown in many different ways. In this section, we will rely on the fact that every 1-random sequence is normal and that a predictor applied to a 1-random sequence produces a 1-random sequence of errors. In the Section 4.2, a different (and perhaps simpler) proof via martingales will be also provided.
We will start by proving that 1-randomness indeed implies normality.

Lemma 78. *If $x \in 2^{\mathbb{N}}$ is 1-random then*

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{2}$$

Proof. Let $\epsilon > 0$ and consider $V = V_1, V_2, \ldots$ such that

$$V_n = [\![\{\sigma \in 2^{<\mathbb{N}} : |\sigma| = n \wedge |\frac{\sum_{i=1}^{n} x_i}{n} - \frac{1}{2}| \geq \epsilon\}]\!]$$

Observe that the sequence V is uniformly Σ^0_1. One simply needs to enumerate all words and check whether the ratio of nonzero bits exceeds the given threshold (which is computable). Moreover, $\sum_{i=1}^{n} x_i/n$ may be seen as an empirical average of n independent coin tosses, while $1/2$ is the expected value of the unbiased toss. Hence, by Hoeffding's inequality, for each n:

$$\lambda(V_n) \leq 2e^{-2n\epsilon^2}.$$

Hence, V is Solovay test. Consequently, the ratio of nonzero bits in 1-random sequence x deviates from $1/2$ more than ϵ only finitely many times. Since ϵ was arbitrary, we may conclude that x is normal. □

Now, we only need to change the previous proof slightly to obtain the following:

Proposition 79. *Fix a 1-random sequence $A \in 2^{\mathbb{N}}$. Let f be a proper predictor. Then $\varsigma(f, A)$ is defined and equal $\frac{1}{2}$.*

Proof. Consider a sequence $V = V_1, V_2, \ldots$ defined by

$$V_n = [\![\{\sigma : |\sigma| = n \wedge |\frac{\#_1\{\sigma_i : 0 < i \leq n \wedge f(\sigma_1^{i-1}) = 1\}}{n} - \frac{1}{2}| \geq \epsilon\}]\!]$$

for some arbitrary $\epsilon > 0$. Since $\lambda(V_n)$ is bounded by a computable sufficiently fastly decreasing function (by Hoeffding's inequality), V is a Solovay test. This test tells us that in the limit the predictor f is half time wrong when predicting a bit of 1-random sequence to be nonzero. By similar reasoning, we conclude that the same is true for the bits predicted to be zero. □

Corollary 80. *Every* 1*-random sequence is Ko stochastic.*

As we already know, the converse is not true. In the following section, the notion of martingales will be introduced. By combined results of Ambos-Spies et al. [3] and Schnorr [34], we will be able to provide an alternative definitions for Ko stochasticity, Church stochasticity and 1-randomness in terms of martingales. Some relations between these notions will become easily provable.

4.2 Unpredictability via martingales

In the standard presentations of algorithmic randomness theory ([31],[12]), *prediction* and *predictability* come into play in a form of martingales or betting strategies. This particular notion of martingale is somewhat different from its probabilistic counterpart, namely the notion of martingale process (as defined in Definition 51)—although these concepts are related, as we are about to see. *Martingale* formalizes the idea of prediction through betting.

A gambler starts with some amount of capital—without any loss of generality we may asume that the initial capital is equal to 1. Again, at each step, information about past observations is available. But instead of guessing the exact outcome of the future observation, the gambler bets some amount of capital on the outcome. This notion is more general and it may be also seen as representing the degree of confidence that the gambler has in his prediction. In fact, the gambler may consider both outcomes to be equiprobable and effectively abstain from prediction by making a zero bet.

As the new bit is unraveled, the gambler either get richer or loses part of his wealth. The evolution of the capital is governed by a simple fairness condition.

Definition 81. *A function* $d : 2^{<\mathbb{N}} \to \mathbb{R}^{\geq 0}$ *is called a martingale if for all* σ*:*

$$d(\sigma) = \frac{d(\sigma 0) + d(\sigma 1)}{2}$$

A martingale d *succeeds on a sequence* x *if*

$$\limsup_n d(x_1^n) = \infty.$$

Suppose that X is a process given by independent unbiased coin tosses i.e., with the probability distribution given by the uniform measure λ. The value $d(\sigma)$ may be interpreted as the amount of capital available after $|\sigma|$ bets, assuming that σ represents the past outcomes of the process. One can observe that every martingale d is equivalent to a martingale process $Y_n = d(X_1^n)$. Indeed, martingales may be seen a special case

of martingale processes. By Doob's convergence theorem, for every martingale, the amount of capital is bounded from above λ-almost surely.

One of the fundamental ideas in algorithmic randomness is to use Doob's theorem as a starting point for defining various notions of randomness and stochasticity. A class $\mathcal{C}$ of martingales will define a class $\mathcal{S}$ of sequences in the following way: a sequence belongs to $\mathcal{S}$ if no martingale from $\mathcal{C}$ makes an unbounded amount of capital on the sequence. As we are about to see, some very natural classes of martingales correspond to classes of sequences defined earlier in the text.

Definition 82. *A martingale d is simple if there exists some $q \in \mathbb{Q}$ with $0 \leq q \leq 1$ such that for every $\sigma \in 2^{<\mathbb{N}}$ and $j \in \{1, 0\}$ we have $d(\sigma j) \in \{d(\sigma), (1-q)d(\sigma), (1+q)d(\sigma)\}$*

In other words, a simple martingale always bet a fraction q of the accumulated capital or nothing. Furthermore, we may also require the martingale to always bet something:

Definition 83. *A martingale d is strict if for every σ and $j \in \{1, 0\}$ we have $d(\sigma) \neq d(\sigma j)$.*

Downey and Hirschfeldt ascribe the following theorem to Ambos-Spies, Mayordomo, Wang and Zheng [3]. However, the result is implicit in the work of Muchnik, Semenov and Uspensky [30]. In fact, the proof presented here follows their line of reasoning.

Theorem 84 (Muchnik, Semenov and Uspensky [30]; Ambos-Spies, Mayordomo, Wang and Zheng [3])**.** *A sequence is Church stochastic if and only if no simple computable martingale succeeds on it.*

Proof. ($\Leftarrow$) Suppose that x is not Church stochastic. We will contruct a simple computable martingale that succeeds on it. Let f be a computable selection function which witnesses that x is not Church stochastic. Let y denote the subsequence selected by f from x. Without loss of generality assume that the frequency of ones in y is larger than the frequency of zeros. Precisely, assume that for some rational $p = 1/2 + \delta$ with $1/2 \geq \delta > 0$, there are infinitely many prefixes y_1^n such that there are at least pn ones in y_1^n.

Let the initial capital of d be equal to 1. Suppose that $d(\sigma)$ is already defined for some $\sigma \in 2^{<\mathbb{N}}$. Let

$$d(\sigma 1) = \begin{cases} d(\sigma 0) & \text{if } f(\sigma) = \text{no} \quad (4.1) \\ (1+\delta)d(\sigma) & \text{otherwise.} \quad (4.2) \end{cases}$$

That this, d bet some amount of capital on 1 if the bit is selected by f or abstain from betting if it is not. Let m_n be the index of the n-th bit selected by f from x. Now, for every n such that the number of ones in y_1^n is at least pn, the following holds

$$d(x_1^{m_n}) \geq \left((1+\delta)^{1/2+\delta}(1-\delta)^{1/2-\delta}\right)^n. \quad (4.3)$$

Now, observe that

$$(1+\delta)^{1/2+\delta}(1-\delta)^{1/2-\delta} > 1.$$

for $0 < \delta < 1/2$. As we have assumed that there are infinitely many n satisfying the condition 4.3, the amount of capital d is unbounded on x. It remains to observe that d is indeed simple and computable.

($\Rightarrow$) Now, suppose that some simple martingale d succeeds on x. Define a selection function f

$$f(\sigma) = \begin{cases} \text{no} & \text{if } d(\sigma 1) = d(\sigma 0) \quad (4.4) \\ \text{yes} & \text{otherwise.} \quad (4.5) \end{cases}$$

Following the reasoning from the first part of the proof, we conclude that f must select an unbalanced subsequence from x. Hence, x is not Church stochastic. □

Theorem 85 (Ambos-Spies, Mayordomo, Wang and Zheng [3])**.** *A sequence is Ko stochastic if and only if no strict and simple computable martingale succeeds on it.*

Proof. This proof is very similar to the previous one.

($\Leftarrow$) For the implication to the left side, suppose that x is not Ko stochastic. Let f be a predictor which witnesses that x is not Ko stochastic. Recall (see Proposition 66) that f effectively acts as two selection functions partitioning x into two subsequences y and z — corresponding to f predicting 1 and 0, respectively. At least one of them is not balanced. Without loss of generality assume that y has density larger than $1/2$. Now, there is such rational $1/2 \geq \delta > 0$ that one of the following strict and simple martingales succeeds on x,

$$d_1(\sigma 1) = \begin{cases} (1+\delta)d(\sigma) & \text{if } f(\sigma) = 1 \quad (4.6) \\ (1-\delta)d(\sigma) & \text{otherwise.} \quad (4.7) \end{cases}$$

$$d_2(\sigma 1) = (1+\delta)d(\sigma).$$

Following the reasoning from the previous proof, we can see that either d_1 or d_2 succeeds on x.

($\Rightarrow$) This implication is analogous. □

On the other hand, the celebrated theorem by Schnorr [34] gives us a martingale definition of 1-randomness.

Theorem 86 (Schnorr)**.** *A sequence ω is 1-random if and only if no c.e. martingale succeeds on ω.*

Proof. See e.g. Theorem 6.3.4 in [12]. □

Corollary 87. *Every Church stochastic sequence is Ko stochastic. Every 1-random sequence is Church stochastic.*

Proof. Simply observe that every simple computable martingale is computably enumerable and that every simple and strict computable martingale is also a simple computable martingale. Thus, the result follows from Theorems 84, 85 and 86. □

4.3 Complexity of unpredictable

Considering various notions of prediction and predictability, one may quickly observe that our ability to predict a sequence is strongly connected to complexity of the sequence as measured by some familiar notions while being completely uninfluenced by others. Intuitively, there should be some trade-offs between simplicity and predictability. A natural notion of *simplicity* (albeit not the only sound) is given by Turing degrees. In this part we study the levels of Turing hierarchy inhabited by Ko stochastic sequences.

Let us start with a well-known observation based on the famous low basis theorem by Jockusch and Soare [20].

Definition 88. *A degree* $\mathbf{a}$ *is low if its Turing jump* $\mathbf{a}'$ *is* $\mathbf{0}'$.

Theorem 89 (Jockusch and Soare [20])**.** *Every nonempty class of sequences defined by a* Π^0_1 *formula contains an element which is of a low degree.*

Then, we can recall a folklore observation based on the fact that the class of 1-random sequences is, in fact, Π^0_1.

Proposition 90 (folklore, see: [12])**.** *There exists a low* 1*-random sequence.*

Corollary 91. *There exists a* Δ^0_2 *Ko stochastic sequence.*

Proof. Every low sequence is Δ^0_2 and every 1-random sequence is Ko stochastic. □

On the other hand, trivially no 1-random sequence may be c.e. Then again, Ko stochasticity is a much weaker notion that 1-randomness. This prompts a following question—are there any Ko stochastic c.e. sequences? We answer this question negatively below.

Theorem 92. *Let* x *be c.e. Suppose that* x *is normal. Then there exists a predictor* g *such that* $\varsigma(g, x) \neq \frac{1}{2}$.

Proof. Let x be a c.e. sequence. Suppose that x is normal. Let z be a naive predictor that always predicts zero, i.e., $z[2^{<\mathbb{N}}] = \{0\}$. Since x is normal, we have $\varsigma(z, x) = \frac{1}{2}$. Since x is c.e. we can easily construct a predictor g such that g is always correct when z is correct but, for some prefixes, g is correct but z is not. This does not lead us to a contradiction as long as g is better than z infrequently—that is, as long as the difference between predictions of g and z is asymptotically negligible. We will try to construct g for which this does not hold.

As a consequence of A being normal, we have

$$\limsup_{n\to\infty} \frac{\#_1(x^n_1)}{n} > 0$$

In other words, for some $\epsilon > 0$ there are infinitely many k such that

$$\frac{\#_1(x^k_1)}{k} > \epsilon.$$

Hence, we can enumerate x and thus obtain a computable approximation $(x_{(1)}, x_{(2)}, \ldots)$ of x such that, for all $s \in \mathbb{N}$, $x_{(s)} \subseteq x$ and, for infinitely many k,

$$\frac{\#_1((x_{(s)})_1^k)}{|k|} > \epsilon.$$

In the subsequent paragraph, we define a computable function $f : \mathbb{N} \to 2^{<\mathbb{N}}$ which, based on $x_{(s)}$, produces strings that approximate subsequent fragments of x in the following sense: $\alpha_{(t)} \subseteq x_1^{|\alpha_{(t)}|}$, where $\alpha_{(t)} := f(0)f(1)\ldots f(t)$. Crucially, we want from f to satisfy, for each $t \in N$:

$$\frac{\#_1(\alpha_{(t)})}{|\alpha_{(t)}|} > \epsilon. \tag{4.8}$$

To calculate $f(t)$, we might want to find stages $s_0 < s_1 < \cdots < s_t$ and values $0 = k_{-1} < k_0 < k_1 < \cdots < k_t$ such that the proportion of ones in ${x_{(s_i)}}_1^{k_i}$ is $> \epsilon$, and output $f(t) = \tau$ such that ${x_{(s_t)}}_1^{k_{t-1}}\tau = {x_{(s_t)}}_1^{k_t}$. However, it might happen that the proportion of ones in ${x_{(s_t)}}_1^{k_t}$ is $> \epsilon$ while (4.8) is not satisfied because $x_{(s_t)}$ has enumerated many elements below k_{t-1} which do not contribute to $f(t)$ but to $\alpha_{(t-1)} = f(0)f(1)\ldots f(t-1)$ and these has been already fixed by that time. The algorithm for computing $f(t)$ presented below takes this caveat into account and adjusts for the new elements that lag behind.

Algorithm for f. To compute $f(t)$, we will use the following variables: c informs us that $f(c)$ is in preparation, k stores the length of $f(0)f(1)\ldots f(c-1)$, s is for stages in $(x_{(s)})$, i counts new elements and subtracts lags. Initially, all variables are set to 0. (†) Set $s := s + 1$. Let $S := S \cup {x_{(s)}}_1^s - {x_{(s-1)}}_1^s$. If $S = \emptyset$, go to (†). Otherwise, set $i := i + \#\{x \in S : x \geq k\} - \#\{x \in S : x < k\}$ (count new elements $\geq k$, and make allowance for new elements that lag behind k). If $i \leq 0$, go to (†). We have $i > 0$. Let $m := \max S + 1$.

Check if there is l such that $k < l \leq m$ and

$$\frac{\#(S_1^l)}{l} > \epsilon. \tag{4.9}$$

If the is such l, check if $c = t$. If so, output $f(t) = S(k)S(k+1)\ldots S(l-1)$. If $c < t$, increase c by 1, set $k := l$ and go to (†). If there is no such l, also go to (†). This ends the description of the algorithm for f.

Now, we can give the algorithm for the predictor g. Let $\sigma \in 2^{<\mathbb{N}}$. Find t such that $|\sigma| < |\alpha_{(t)}|$ and output $\alpha_{(t)|\sigma|}$.

Now, suppose that g acts on prefixes of x. f was constructed is such a way that it gives us approximations of prefixes of x and by this, for every i, if g predicts x_i to be nonzero, then x_i is indeed nonzero. On every other bit, g is correct if and only if z would be. Furthermore, we know that there are infinitely many k such that

$$\left|\varsigma(g, x_1^k) - \varsigma(z, x_1^k)\right| > \epsilon$$

where

$$\lim_{k\to\infty} \varsigma(z, x_1^k) = \frac{1}{2}.$$

Thus, we can conclude that

$$\varsigma(g, x) \neq \frac{1}{2}.$$

□

Corollary 93. *No c.e. set is Ko stochastic.*

Proof. Suppose that A is c.e. and Ko stochastic. Consider a naive predictor z that always place zero as the prediction, i.e., $z[2^{<\mathbb{N}}] = \{0\}$. Since A is Ko stochastic, for every f we have $\varsigma(f, A) = \frac{1}{2}$. In particular, this is true for z, i.e., $\varsigma(z, A) = \frac{1}{2}$. This means that A is normal. By Theorem 92, there is a predictor g such that $\varsigma(g, A) \neq \frac{1}{2}$ and so, A cannot be Ko stochastic. □

4.3.1 n-c.e. sequences

A natural generalization of computable enumerability is given by the so-called Ershov hierarchy [13][33]. We start by recalling a definitions of ω-c.e. and n-c.e. sets. ω-c.e. sequences are such that may be computably approximated and for each bit the number of "changes of mind" in the approximation is bounded by a computable function. A sequence will be called n-c.e. if this computable function is constant.

Definition 94. *A sequence A is ω-c.e. if there is a computable collection of sequences $\{A_{(s)}\}_{s\in\mathbb{N}}$ with $A_0 = \emptyset$ and a computable function g such that, for all i*

$$A_i = \lim_{s\to\infty} (A_{(s)})_i \wedge \#\{s : (A_{(s)})_i \neq (A_{(s+1)})_i\} \leq g(i).$$

We will say that A is n-c.e. if and only if $g(i) \leq n$, for all i. The value $\#\{s : (A_{(s)})_i \neq (A_{(s+1)})_i\}$ is called the number of mind changes in the approximation on the i-th bit.

n-c.e. sequences may be also defined as finite combinations of unions and differences of c.e. sequences.

Lemma 95. *Let C be a 2-c.e. set. There exist A and B such that A and B are c.e. and $C = A - B$.*

Proof. Consult e.g. section 3.8.4 of [35]. □

The relation between the Ershov hierarchy and 1-randomness was studied, for example, by Figueira et al. [14].

Proposition 96 ([14])**.** *There exists an ω-c.e. 1-random sequence.*

Corollary 97. *There exists an ω-c.e. Ko-stochastic sequence.*

The results by Figueira et al. [14] show that no 1-random sequence may be n-c.e. Here, we give a similar observation for Ko stochasticity.

Problem 98. *Are there any n-c.e. Ko stochastic sequences?*

We will show that no n-c.e. sequence is Ko stochastic. To this end we generalize some ideas contained in Theorem 92. This generalization is contained in the subsequent lemma.

Lemma 99. *Let $n > 0$. Suppose that $(B_{(s)})$ is a computable approximation to B such that, for all $x \in \mathbb{N}$, $\{x : \exists^{\geq n} s\ (B_{(s)})_x \neq (B_{(s+1)})_x\}$ has density $\neq 0$. Then there is a predictor g such that for every k, if $g(B^k_1) = 1$, then there is at least n mind changes in the approximation on the k-th bit. Moreover, the density of bits on which this happens is nonzero.*

Proof. We only provide an outline of the proof. The main idea is similar to the one used in constructing the predictor g in the proof of Theorem 92 for a c.e. set A having density $\neq 0$. The main difference is that, in place of such A, we consider elements for which $(B_{(s)})$ makes at least n mind changes (note that this set is c.e.). At any given stage s of the algorithm computing $f(t)$, we put such new elements into S. We then proceed without changes. Having f, we define g as earlier. □

Theorem 100. *No n-c.e. sequence is Ko stochastic.*

Proof. We proceed by induction on n. Corollary 93 proves this statement for $n = 1$. Moreover, by Corollary 93, for each 1-c.e. sequence A there is a predictor h such that $\varsigma(h, A) \neq 1/2$ and h depends only on the length of the input (i.e., for every $\sigma, \tau \in 2^{<\mathbb{N}}$, if $|\sigma| = |\tau|$ then $h(\sigma) = h(\tau)$).

Let $n > 0$ and assume that, for each n-c.e. sequence A, there exists a predictor h such that $\varsigma(h, A) \neq 1/2$ and h depends only on the length of the input. It is sufficient to show that a similar predictor exists for every $(n + 1)$-c.e. sequence.

Fix an $(n + 1)$-c.e. sequence B with a computable approximation $(B_{(s)})$. We ask about the density of the set of bits on which there are exactly $n + 1$ mind changes in $(B_{(s)})$. We consider two cases: when this set is of density zero or otherwise.

Suppose that the set of bits on which there are exactly $n + 1$ mind changes in the approximation $(B_{(s)})$ is of density zero. Take a computable approximation $(A_{(s)})$ which is exactly as $(B_{(s)})$ except that it ignores each $(n + 1)$th mind change and keeps its previous answer. $(A_{(s)})$ is a computable approximation of an n-c.e. sequence, denote it by A. By the inductive assumption, there exists a prefix independent predictor h such that $\varsigma(h, A) \neq 1/2$. Since h is prefix independent, we have $h(A^i_1) = h(B^i_1)$, for all $i \in \mathbb{N}$. But $\varsigma(h, B) \neq 1/2$ must hold as well, because A and B differ at the set of density zero. Hence, B is not Ko stochastic.

Now, suppose that the density of bits with exactly $n + 1$ mind changes in $(B_{(s)})$ is nonzero. Let g be the predictor from Lemma 1 defined for $n+1$ mind changes in $(B_{(s)})$. Consider two predictors, f_n and f_{n+1}, such that

$$f_n(\sigma) = 1 - b \iff g(\sigma) = 1,$$

$$f_{n+1}(\sigma) = b \iff g(\sigma) = 1.$$

where $b = 1$ if n is even and $b = 0$ otherwise.
Observe that, for every i,

$$g(B^i_1) = 1 \implies f_n(B^i_1) \neq B_i \wedge f_{n+1}(B^i_1) = B_i.$$

Since the density of bits with $n+1$ changes is not zero (and thus the density of bits i such that $g(B_1^i) = 1$ is not zero) we may conclude that, for some $\epsilon > 0$, there are infinitely many k such that

$$|\varsigma(f_n, B_1^k) - \varsigma(f_{n+1}, B_1^k)| > \epsilon.$$

Consequently, it is not possible that $\varsigma(f_n, B) = 1/2$ and $\varsigma(f_{n+1}, B) = 1/2$ at the same time. Therefore, B is not Ko stochastic. Moreover, both f_n and f_{n+1} depend only on the length of the input.

□

5

Unstable prediction

We have defined the prediction error of f on the infinite sequence A as the limit of ratio of correct and wrong answers on the initial segments of A. In some cases such limit exists, in some it does not. For the sake of notational simplicity we introduce the following:

Definition 101 (error stability)**.** *A sequence $2^{\mathbb{N}}$ is error stable if there is a predictor f such that $\varsigma(f, A)$ exists. If the sequence is not error stable then we say it is error unstable.*

There are many examples of error stable sequences. For example, every computable sequence is necessarily stable. Indeed, every computable sequence has a perfect predictor. On the other hand, stability does not imply that the sequence is effectively predictable. We have already mentioned the natural notion of unpredictability in this framework, namely, Ko stochasticity. Every Ko stochastic sequence is necessarily error stable. However, not every sequence is error stable. In the sequel of this section, we consider several methods for obtaining error unstable sequences.

5.1 Weak genericity

An easy way to obtain error unstable sequence is via weak genericity. We begin with Proposition 103 which states that every weakly generic set is not error stable.[1] But first some notation is in order. Given $S \subseteq 2^{<\mathbb{N}}$ and $A \in 2^{\mathbb{N}}$, we say that A meets S if there is an $n > 0$ such that $A_1^n \in S$.

Definition 102. *$A \in 2^{\mathbb{N}}$ is weakly generic if for every nonempty Σ_1^0 set S of words, A meets S.*

The existence of weak generics might be established by a straightforward construction, recursive in $0'$. The notion of (weak) genericity has been studied extensively, see [25], and [12] for an overview.

[1] This was suggested to us by an anonymous referee of an early version of [23]

Proposition 103. *Let $A \in 2^{\mathbb{N}}$ be weakly generic. Then for every predictor f, the prediction error $\varsigma(f, A)$ is undefined.*

Proof. Let $A \in 2^{\mathbb{N}}$ be a weakly generic set and let f be a predictor. For the sake of contradiction, suppose that $\varsigma(f, A)$ exists and let $p = \varsigma(f, A)$. Without loss of generality, assume that $0 < p \leq \frac{1}{2}$. Fix $\epsilon > 0$ and k such that for every $n > k$, $\varsigma(f, A^n_1) < \frac{1}{2} + \epsilon$. Let $S \subseteq 2^{<\mathbb{N}}$ be the set of words σ such that $|\sigma| > k$ and $\varsigma(f, \sigma) \geq \frac{1}{2} + \epsilon$. This guarantees that no prefix A^n_1 is in S. So, A does not meet S.
Obviously, S is a Σ^0_1 set of words. Since A does not meet S, A is not weakly generic—a contradiction. □

As a corollary we get a strengthened version of the known separation between weak genericity and Martin-Löf randomness (since every Martin-Löf random sequence is Ko stochastic).

Corollary 104. *No weakly generic sequence is Ko stochastic.*

5.2 Unstable prediction in c.e. degrees

By the result reported in [12] (Proposition 2.24.2), every noncomputable c.e. set computes a weakly generic set. Hence, by Proposition 103, for each c.e. set $B >_T 0$ there is a computable in B sequence A that is not error stable.

Can a set be c.e. and not error stable? This question cannot be settled by further appeal to weak genericity, since no weak generic set is c.e. Nonetheless, in Theorem 106, we provide a construction of a c.e. set that is not error stable. We start with a construction in $0'$ (Theorem 105) which shows how to construct a sequence that is not error stable using requirements based solely on predictors. Theorem 106 is essentially an effectivization of this idea, drawing heavily on the finite-injury priority method [16, 29].

Theorem 105. *There is a sequence A that is not error stable and $\mathbf{0} <_T A \leq_T \mathbf{0}'$.*

Informal description. We want to construct a sequence A such that for every predictor f, $\varsigma(f, A)$ is undefined. We use an effective enumeration of all partial computable functions from $2^{<\mathbb{N}}$ to $\{0, 1\}$ (such an enumeration contains all predictors). We proceed with the finite extension method. The key idea is to take care of each predictor at infinitely many stages and at each such stage force its prediction error to go sufficiently up or down—this will prevent the predictor from having a defined prediction error on the constructed sequence. As we shall see, it suffices to ensure that for every predictor the prediction error goes below some fixed level infinitely many times. The essential idea is that each predictor occurs infinitely often in the effective canonical enumeration of partial computable functions.

Proof. Fix a canonical enumeration of all partial computable functions $\Phi_0, \Phi_1, \Phi_2, \ldots$ from $2^{<\mathbb{N}}$ to $\{0, 1\}$. The construction proceeds by stages. We start with $\alpha_{(0)} = \emptyset$. At each stage $s + 1$ we construct a finite sequence $\alpha_{(s+1)} \succeq \alpha_{(s)}$. The constructed sequence is defined by $A = \bigcup_{s=0}^{\infty} \alpha_{(s)}$.

Construction

Stage 0: $\alpha_{(0)} = \emptyset$.

Stage s+1: Check whether there exists $\tau \neq \square$ such that for every $\rho \prec \alpha_{(s)}\tau$ we have $\Phi_s(\rho)\downarrow$, and

$$\varsigma(\Phi_s, \alpha_{(s)}\tau) < \frac{1}{4}.$$

Now, if such τ do exist, let $\tau_{(0)}$ be the least one and set $\alpha_{(s+1)} = \alpha_s\tau_{(0)}$. If such τ does not exist, set $\alpha_{(s+1)} = \alpha_{(s)}$. This ends the construction.

Verification First, observe that A is infinite. For a contradiction, let $A = \alpha$, for some $\alpha \in 2^{<\mathbb{N}}$. Let t be the least stage such that, for all $s \geq t$, $\alpha_{(s)} = \alpha$. Write a program with an index u satisfying $\varsigma(\Phi_u, \alpha 0) = 0$ and $u \geq t$. But then $\alpha_{(u+1)} = \alpha_{(u)}0 \neq \alpha$, a contradiction.

Now, we show that no predictor has a defined prediction error for A. Let f be a predictor. By totality of f, at each stage $s+1$ with $\Phi_s = f$, α_s is extended to some $\alpha_{(s+1)}$ satisfying

$$\varsigma(\Phi_s, \alpha_{(s+1)}) < \frac{1}{4}.$$

(For example, we can obtain such an extension as follows: set $b_1 = 1 - f(\alpha_{(s)})$, $b_2 = 1 - f(\alpha_s b_1)$, ..., $b_k = 1 - f(\alpha_{(s)}b_1b_2...b_{k-1})$. For sufficiently large k, we will have $\varsigma(\Phi_s, \alpha_{(s)}\tau) < 1/4$, where $\tau = b_1b_2...b_k$.) Now, f occurs infinitely often in the enumeration $\Phi_0, \Phi_1, \Phi_2, \ldots$. Therefore, our construction ensures that for every $k \in \mathbb{N}$ there is an $n > k$ such that $\varsigma(f, A_1^n) < \frac{1}{4}$.

On the other hand, let f' be the predictor defined by $f'(\sigma) = 1 - f(\sigma)$, for all $\sigma \in 2^{<\mathbb{N}}$. Note that $\varsigma(f', A_1^n) < 1/4$ for infinitely many n. Observe that for every $n \in \mathbb{N}$ we have $\varsigma(f, A_1^n) = 1 - \varsigma(f', A_1^n)$. This, in turn, guarantees that for every $k \in \mathbb{N}$ there is an $n > k$ such that $\varsigma(f, A_1^n) > \frac{3}{4}$. Consequently, $\varsigma(f, A)$ does not exist, as required.

To show that $\mathbf{0} <_T A$, it suffices to observe that every computable sequence has a predictor with a defined prediction error. Obviously, $A \leq_T \mathbf{0}'$: to compute A_n in $\mathbf{0}'$, perform the construction up to the first stage s satisfying $|\alpha_{(s)}| > n$ and return $(\alpha_{(s)})_n$. $\square$

Theorem 106. *There exists a noncomputable c.e. set that is not error stable.*

In the proof of Theorem 105 we attempted to satisfy, for each $e \in \mathbb{N}$, the following requirement (denote it by R_e): if Φ_e is total then there exists n such that $\varsigma(\Phi_e, A_1^n) < p$ (for $p = \frac{1}{4}$). Now, we want to satisfy each R_e in an effective way. We define a computable approximation $(A_{(s)})_{s\in\mathbb{N}}$ with $\lim_{s\to\infty} A_{(s)} = A$. To ensure that A is c.e., we make sure that we always have $A_{(s)} \subseteq A_{(s+1)}$, for all $s \in \mathbb{N}$.

Strategy in isolation Suppose we are at stage s of the construction and we start to care about satisfying R_e. We choose a fresh starting point n_e (*fresh* means, among other things, that $(A_{(s)})_n = 0$ for $n \geq n_e$), big enough number $f(n_e)$ and we initialize a variable h_e, called head, with n_e. At every subsequent stage $t > s$, we perform the

following steps: if $h_e \leq f(n_e)$ then we check whether $\Phi_{e,t}((A_{(t)})_1^{h_e}) \downarrow$ holds, and if it holds, we set $(A_{(t)})_{h_e} = \Phi_e(A_1^{h_e})$ and move the head to the right, i.e. we set $h_e = h_e + 1$.

The above description states how to satisfy a single requirement in isolation. If Φ_e is total then, eventually, Φ_e will make correct predictions $\Phi_e(A_1^k) = A_k$ for $k \in [n_e, f(n_e)]$, and thus $\varsigma(\Phi_e, A_1^{f(n_e)}) < p$, provided $f(n_e)$ is sufficiently large.

Strategies together When multiple requirements act together, some conflicts may appear. Suppose we have a requirement R_i which we try to satisfy on $[n_i, f(n_i)]$ and there is a different requirement R_j which we try to satisfy on $[n_j, f(n_j)]$ with $f(n_i) < n_j$ (in general, we will always make sure that such intervals are disjoint). R_i may want to put some number $k \in [n_i, f(n_i)]$ into A. But this may destroy computations made earlier for R_j (for example, we had $\Phi_j(A_1^{k'}) = 1$ for some $k' \in [n_j, f(n_j)]$ when k was not in A, but now, when k is in A, $\Phi_j(A_1^{k'}) = 0$). In such a case, we simply reset all requirements $j > i$, i.e. we set their heads h_j back to their corresponding starting points n_j. It is possible that such R_j will succeed on its current segment $[n_j, f(n_j)]$.

It may happen, however, that R_i cannot be satisfied on its current segment. This occurs when a number k has already been enumerated into A but later in the construction, during our attempts to satisfy R_i, it turns out that $\Phi_i(A_1^k) = 0$ and so we should withdraw k from A. We cannot let it happen as we want $(A_{(s)})_{s\in\mathbb{N}}$ to be a non-decreasing sequence of sets. Therefore, we stop our attempts to satisfy R_j on $[n_j, f(n_j)]$, for $j \geq i$, and try to find new values for n_j and $f(n_j)$ in the future.

Construction Let $\Phi_0, \Phi_1, \ldots$ be a fixed effective enumeration of all partial computable functions from $2^{<\mathbb{N}}$ to $\{0, 1\}$. Let $0 < p < 1$ be such that checking whether $x < p$ is computable. We construct A such that for all e:

$$\Phi_e \text{ is total } \implies \varsigma(\Phi_e, A_1^k) < p \qquad (R_e)$$

$R_0 > R_1 > R_2 > \ldots$ is the priority ordering. For each R_e we have *variables* n_e, h_e and $\varphi(e)$ for storing, respectively, the current starting point, head and constraint of R_e. Contents of these variables may change during construction. We define $f(n) = n + \min\{k \in \mathbb{N} : \frac{n}{n+k} < p\}$.

R_i will make changes to certain segments of the sequence under construction, beginning with a starting point n_i, and ending with $f(n_i)$. h_i will be used to navigate through bits $n_i, n_i + 1, \ldots, f(n_i)$.

We say that R_e requires attention at stage $s+1$ if n_e is defined at stage s, $h_e \leq f(n_e)$ and $\Phi_{e,s}((A_{(s)})_1^{h_e}) \downarrow$.

Stage 0. $A_{(0)} = \emptyset$. All $h_e, n_e, \varphi(e)$ are undefined.

Stage $s + 1$. If no requirement needs our attention, let R_j be the highest-priority requirement with n_j being undefined. Let n_0 be the new fresh number, i.e. the least number m such that: $\forall r \geq m\, (A_{(s)})_r = 0$ and $m >$ every currently defined constraint $\varphi(e)$ (at each stage, only finitely many values of φ are defined). Reserve n_0 as the starting point for R_j (i.e. set $n_j = n_0$), put a constraint $\varphi(j) = f(n_j)$ and set $h_j = n_j$.

Proceed to the next stage. In simple words, we chose n_0 so that the greatest integer enumerated so far into A is $< n_0$; moreover, we want R_j to operate on a segment that is separate from the segments attached to other the requirements.

If some requirement needs our attention, let R_i be the one with the highest priority. Suppose $\Phi_{i,s}((A_{(s)})_1^{h_i}) = 0$. If $(A_{(s)})_{h_i} = 1$ then cancel all requirements R_j with $j \geq i$ (it means that $n_j, h_j, \varphi(j)$ become undefined). If $(A_{(s)})_{h_i} = 0$ then set $h_i = h_i + 1$. Now, suppose $\Phi_{i,s}((A_{(s)})_1^{h_i}) = 1$. If $(A_{(s)})_{h_i} = 1$ then set $h_i = h_i + 1$. If $(A_{(s)})_{h_i} = 0$ then set $(A_{(s+1)})_{h_i} = 1$, $h_i = h_i + 1$ and reset all requirements j with $j > i$ (i.e., make $h_j = n_j$ again).

Lemma 107. *Every requirement receives attention only finitely often.*

Proof. Let $i \in \mathbb{N}$ and assume that all requirements $j < i$ receive attention only finitely often. Let s be the least stage such that: R_i has an attached starting point $n_i = n$ at this stage and R_j does not receive attention at stages $t \geq s$, for all $j < i$. It means that at these stages R_i cannot be canceled or reset by any other requirement.

Observe that R_i operates on at most two different segments at stages $t \geq s$. At first, it operates on $[n, f(n)]$. Suppose that R_i receives attention at some stage $t \geq s$ when operating on $[n, f(n)]$ and the conditions $\Phi_{i,t}((A_{(t)})_1^{h_i}) = 0$, $(A_{(t)})_{h_i} = 1$ are not satisfied at the same time. Given that, R_i receives attention at most $f(n) - n$ times while being attached to $n_i = n$. This is because R_i starts operating on $[n, f(n)]$ with $h_i = n$ and whenever we see $\Phi_{i,u}((A_{(u)})_1^{h_i}) \downarrow$, we possibly modify A_{h_i} (by making $(A_{(s+1)})_{h_i} = 1$) and we advance h_i. But from now on, for the current value of h_i, $A_1^{h_i}$ remains unchangeable, i.e. $A_1^{h_i} = (A_{(u+1)})_{h_i}$. Hence, either h_i gets stuck because of a divergent computation of Φ_i for some $A_1^{h_i}$, or it eventually reaches $f(n) + 1$ and from that stage on we always have $h_i > f(n)$. In either case, R_i never receives attention anymore which means that R_i receives attention at most $f(n) - n$ times while being attached to $n_i = n$.

Now, suppose that at some stage $t \geq s$ the requirement R_i receives attention and we have $\Phi_{i,t}((A_{(t)})_1^{h_i}) = 0$ and $(A_{(t)})_{h_i} = 1$. Let r be the number of times R_i has received attention while being attached to $n_i = n$. According to the construction, R_j with $j \geq i$ are cancelled. Hence, R_i is cancelled itself and gets new values $n_i = n'$, $h_i = n'$, $\varphi(i) = f(n')$ at the next stage $t + 1$. Observe that $[n', f(n')]$ is a completely fresh segment with no numbers enumerated so far into A. Hence, if R_i receives attention at some later stage $u > t + 1$, we cannot have $(A_{(u)})_{h_i} = 1$ and hence we cannot have $\Phi_{i,u}((A_{(u)})_1^{h_i}) = 0$ and $(A_{(u)})_{h_i} = 1$ simultaneously. The reasoning from the previous paragraph convinces us that R_i receives attention at most $f(n')$ times while being attached to $n_i = n'$. Hence, R_i has received attention at most $r + f(n') - n'$ times while being attached first to n and then to n'.

□

Lemma 108. *Every requirement is eventually satisfied.*

Proof. Let Φ_e be total. By Lemma 107, as of a certain stage, R_e is never injured by other requirements and eventually operates on at most two final segments. Since Φ_e is total, the head h_e gets updated until it reaches $f(n) + 1$ where $[n, f(n)]$ is one of these

segments. Hence, we have $A_{m+1} = \Phi_e(A_1^m)$ for $m \in [n, f(n)]$. By the definition of f, we have

$$\varsigma(\Phi_e, A_1^{f(n)}) \leq \frac{n}{f(n)} < p,$$

which is as desired. □

The proof of Theorem 105 shows that Lemma 108 is sufficient to guarantee that $\varsigma(\Phi_e, A)$ is undefined for all total Φ_e. This completes the proof of Theorem 106.

6

Optimality

Intuitively, some sequences can be better predicted then others. Hence, we might want to compare them with respect to the level of their predictability. How this intuition could be made precise?

Suppose we have sequences A and B. We want to know if we can predict A *better* than we can predict B. One idea would be to look whether the best possible predictor for A is *better* than the best predictor for B. Apparently, this is the approach taken by C.W. Granger in his seminal paper on causality tests where he measures the predictability in terms of the error series of an *optimum, unbiased, least-squares predictor* [17].

It comes as no surprise that the existence of an optimal predictor is not guaranteed in every case. As we will see, in case of binary sequences, there exists a sequence with no optimal predictor. Furthermore, the existence of an optimal predictor turns out not to be so important—as we can have infinite progressions of better and better predictors approaching some level of predictability but never attaining it. Therefore, the comparison of optimal predictors does not form a good basis for the comparison of the sequences' predictability.

We start with the formal definition and then proceed to present the results.

Definition 109 (optimal predictor)**.** *f is called an optimal predictor for S if $\varsigma(f, S)$ is defined and for every proper predictor g we have:*

$$\varsigma(f, S) \leq \varsigma(g, S)$$

or $\varsigma(g, S)$ is not defined.

Proposition 110. *Fix $A \in 2^{\mathbb{N}}$ and let f be an optimal predictor for A. Suppose that $\varsigma(f, A) = \frac{1}{2}$. Then, for every predictor g for A, if $\varsigma(g, A)$ is defined, then $\varsigma(g, A) = \frac{1}{2}$.*

Proof. This is a corollary of Proposition 62. □

6.1 Optimal predictors with an uncomputable prediction error

Theorem 111. *For every real number $0 \leq r \leq \frac{1}{2}$ there exists a sequence A and a predictor f such that f is optimal for A and $\varsigma(f, A) = r$.*

Proof. Fix a real number $0 \leq r \leq \frac{1}{2}$. Let $I \in 2^{\mathbb{N}}$ be such that

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} I_i}{n} = 2r. \tag{6.1}$$

We use I to construct A. We want to put $A_i = 0$ for every i such that $I_i = 0$. We know that there are $1-2r$ such bits (in the limit). As for the subsequence of A corresponding to the nonzero bits of I, we want it to be unpredictable in the sense that for every predictor f

$$\lim_{n\to\infty} \frac{\#\{0 \leq i < n-1 : A_{i+1} \neq f(A_1^i) \wedge I_{i+1} = 1\}}{n} = \frac{1}{2}.$$

This can be done using Theorem 67. Each predictor g gives rise to two selection functions: g itself and $\overline{g} = 1 - g$.[1] Note that if g and $\overline{g}$ both select a balanced subsequence on some sequence Y, then $\varsigma(g, Y) = \frac{1}{2}$.

We want to have an unpredictable subsequence corresponding to the nonzero bits from I. Hence, for each g, we consider $g^{'}$ and $\overline{g}^{'}$ such that for all $\sigma \in 2^{<\mathbb{N}}$:

$$g^{'}(\sigma) = g(\sigma) I_{|\sigma|+1} \tag{6.2}$$

$$\overline{g}^{'}(\sigma) = \overline{g}(\sigma) I_{|\sigma|+1} \tag{6.3}$$

Now, let $g_0, g_1, \ldots$ be a complete list of all proper predictors. Consider a countable collection of selection functions $\mathcal{G} = g_0^{'}, \overline{g}_0^{'}, g_1^{'}, \overline{g}_1^{'}, \ldots$. By Theorem 67, there is a sequence A such that every function from $\mathcal{G}$ is balanced on A. Moreover, $I_i = 0$ implies $A_i = 0$ for every $i \in \mathbb{N}$.

Now, consider a predictor f which always predicts zero. Obviously, it correctly predicts every bit of the subsequence of A corresponding to zero bits of I. As for the rest of A (where I has ones), the construction guarantees that the relative limiting frequency of zeros (and ones) equals one half on the places selected by $\overline{f}^{'}$. But these are precisely all nonzero bits of I. Hence, in the limit, f erroneously predicts half of these bits. Since the frequency of nonzero bits of I approaches $2r$ (see Equation (6.1)), the frequency of errors made by f on A approaches r, i.e. $\varsigma(f, A) = r$.

It remains to observe that no other predictor is better. It follows from the fact that all predictors are equally good on bits selected by $\overline{f}^{'}$ but none is better than f on the rest of bits (which are all zeroes).

□

[1] This observation is due to an anonymous reviewer of an early version of the paper [23].

Table 6.1: The assigment of bits X_i to predictors h_k

$h_1 \longrightarrow$	·	·	·	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}	·	·	…			
⋮																			

6.2 Sequences with no optimal predictor

By the fundamental theorem of arithmetic we notice the following fact.

Lemma 112. *For every even number $k > 0$ there exist natural $m \geq 1$ and $n \geq 0$ such that*

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

Every number can be uniquely represented by such m and n. Moreover, k is odd if and only if $m = 1$.

Theorem 113. *There exists $X \in 2^{\mathbb{N}}$ and an infinite sequence of proper predictors $f_1, f_2, \ldots$ such that for all $n \in \mathbb{N}$*

$$\varsigma(f_n, X) > \varsigma(f_{n+1}, X),$$

but no predictor is optimal for X.

Proof. Let $h_0, h_1, \ldots$ be a listing of all proper predictors. We construct X inductively. Set $X_1 = 0$. Let $i > 1$ and assume that X_j has been defined for $j < i$. By Lemma 112, let n, k be the unique integers such that $i = (1/2 + n)2^{k+1}$. We set

$$X_i = 1 - h_k(X_1^{i-1}). \tag{6.4}$$

Table 6.1 shows in a visual way how bits of X are assigned to the predictors that are used to define them.
First, we show that there is a proper predictor f such that $\varsigma(f, X)$ exists. We define total computable $f : 2^{<\mathbb{N}} \to \{0, 1\}$ as follows:

$$f(\sigma) = \begin{cases} 0 & \text{if } |\sigma| \text{ is even} & (6.5) \\ 1 - h_0(\sigma) & \text{if } |\sigma| \text{ is odd and } \sigma \text{ ends with 1} & (6.6) \\ h_0(\sigma) & \text{if } |\sigma| \text{ is odd and } \sigma \text{ ends with 0} & (6.7) \end{cases}$$

We prove the following claim: for every $n \in \mathbb{N}$, if n is even, then *either* $f(X_1^n)$ *or* $f(X_1^{n+1})$ is correct.

To see this, let n be even. We have two cases depending on whether $f(X_1^n)$ is correct. First, assume that $f(X_1^n)$ is correct, i.e. $X_{n+1} = f(X_1^n)$. We want to show that $f(X_1^{n+1})$ is incorrect. $|X_1^n|$ is even so, by (6.5), $f(X_1^n) = 0$. Consequently, $X_{n+1} = 0$. Hence, X_1^{n+1} ends with 0. $|X_1^{n+1}|$ is odd and ends with 0, so, by (6.7) and (6.4), $f(X_1^{n+1}) = h_0(X_1^{n+1}) = 1 - X_{n+2}$ which means that $f(X_1^{n+1})$ is incorrect.

As for the second case, we assume that $f(X_1^n)$ is incorrect and show that $f(X_1^{n+1})$ is correct. The reasoning is similar, except that we obtain $X(_n= 1$ and apply case (6.6).

Having the above claim proven, it is routine to see that

$$\varsigma(f, X) = \frac{1}{2}.$$

Now, suppose that g is a proper predictor and $\varsigma(g, X)$ is defined. Since g is a proper predictor, there is an i such that $h_i = g$. Choose such an i. By the definition of X, g makes mistakes on every bit X_j such that $j = (1/2 + k)2^{i+1}$ for some k. Consider a predictor h having the following property: for every such bit j, $h(j) = 1 - g(j)$ and for any other bit m, $h(m) = g(m)$. Therefore,

$$\varsigma(h, X) = \varsigma(g, X) - \frac{1}{2^{i+1}}.$$

This means that h is better than g on X.

From the above paragraph it follows easily that no proper predictor is optimal for X. It remains to show that there is an infinite sequence $f_1, f_2, \ldots$ of increasingly better predictors for X. Start with $f_1 = f$. By the earlier considerations, f_1 has a defined prediction error. Given f_n with a defined $\varsigma(f_n, X)$, obtain f_{n+1} from f_n by an application of the procedure described in the previous paragraph. This yields $\varsigma(f_n, X) > \varsigma(f_{n+1}, X)$. This ends the proof. □

We can combine proofs of Theorem 111 and Theorem 113 to get the following:

Corollary 114. *Let $p \in \mathbb{R}$ be such that $0 < p < \frac{1}{2}$ and*

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} P_i}{n} = p,$$

for some computable sequence P. Then there exists a sequence $A \in 2^{\mathbb{N}}$ and an infinite sequence of predictors $f_1, f_2, \ldots$ such that

$$\lim_{n\to\infty} \varsigma(f_n, A) = p,$$

and for every predictor g with a defined prediction error, $\varsigma(g, A) > p$.

Proof. Fix a number p which satisfies the antecedent of the implication and let P be such that

$$\lim_{n\to\infty} \frac{\sum_{i}^{n-1} P_i}{n} = 2p.$$

The proof is then straightforward. We construct a sequence A recursively. On every i-th bit of A such that $P_i = 0$ we will place bits in similar manner as in the proof of Theorem 113. That is, every second such bit will correspond to some predictor h_0, then every second of the rest corresponds to predictor h_1 and so on (where $h_0, h_1, \ldots$ is some fixed canonical enumeration of predictors). Again, we want the subsequence of

A corresponding to nonzero bits of P to be unpredictable. That is, we ensure that for every predictor f

$$\lim_{n\to\infty} \frac{\#\{0 \leq i < n-1 : A_{i+1} \neq f(A_1^i) \wedge P_{i+1} = 1\}}{n} = \frac{1}{2}.$$

Combining this with the argument from the proof of Theorem 113 we get the consequent of the implication. □ □

Observe that we have to be able to approximate p computably. Otherwise, we could not compute P and the diagonal argument of Theorem 113 would fail. This raises the following question:

Problem 115. *Is the consequent of the implication from Corollary 114 true for every real number p such that $0 < p < 1/2$?*

7

Related approaches

7.1 Tadaki predictability

A different notion of 'predictability' was proposed by K. Tadaki [36]. He considered a slightly different model, where predictors may sometimes abstain from giving an answer. This is similar to martingales which in general may bet zero capital on their guesses (thus, effectively the answer does not matter). On the other hand, Tadaki is interested only in such prediction where either we are correct or the predictions are suspended. The notion introduced below is denoted as *total strong predictability* in [36]. Here, we will use the name *Tadaki-predictability*. Note that in [36] both prediction is restrained by either (total) computable or, more broadly, partial computable functions. Here, we will only deal with the computable prediction in Tadaki sense.

Definition 116 (T-predictor)**.** *A predictor in the Tadaki sense or simply — a T-predictor — is a total computable function* $2^{<\mathbb{N}} \to \{0, 1, \square\}$.

The $\square$ here is interpreted as a suspension of the prediction.

Definition 117 (Tadaki Strong Predictability)**.** *Let* $A \in 2^{\mathbb{N}}$. *We say that* A *is Tadaki-predictable if there exists a T-predictor* f *for which the following conditions hold:*

1. *For every* $n \in \mathbb{N}$, *if* $f(A^n_1) \neq \square$ *then* $f(A^n_1) = A_{n+1}$.
2. *The set* $\{i \in \mathbb{N} : f(A^i_1) \neq \square\}$ *is infinite.*

In [36] Tadaki studied the relation between such notion of 'predictability' and various classes of randomness. In particular, he observed that computable randomness implies that a sequence is not Tadaki-predictable. Recall the definition of computable randomness.

Definition 118 (Computable Randomness)**.** *A sequence* A *is computably random if and only if there is no computable martingale succeeds on* A.

Without much effort we get this simple observation

Proposition 119 (Tadaki [36])**.** *If $A \in 2^{\mathbb{N}}$ is computably random then it is not Tadaki-predictable.*

Tadaki proves it by contradiction via a direct construction of a succeeding martingale. However, Proposition 119 becomes self-evident if we recall the Theorem 84.

Proposition 120. *Every computably random sequence is Church stochastic.*

Proof. By Theorem 84 a sequence is Church stochastic if and only if no simple computable martingale succeeds on it. A sequence is computably random if no computable martingale succeeds on it. In particular, if the sequence is computably random, then no simple computable martingale succeeds on it. □

Now, it remains to observe that

Proposition 121. *No Church stochastic sequence is Tadaki predictable.*

Proof. Suppose that A is Church stochastic and Tadaki predictable. There is a T-predictor f such that

1. For every $n \in \mathbb{N}$, if $f(A^n_1) \neq \square$ then $f(A^n_1) = A_{n+1}$.
2. The set $\{i \in \mathbb{N} : f(A^i_1) \neq \square\}$ is infinite.

Fix such an f and consider a selection function g such that for all $\sigma \in 2^{<\mathbb{N}}$ the following holds:

$$f(\sigma) = 1 \Leftrightarrow g(\sigma) = 1.$$

Observe that g is not balanced on A and conclude that A is not Church stochastic. □

Now we turn our relation to the interplay between the Tadaki predictability and prediction based on the predictors. The notion of the Tadaki predictability may be seen as being 'orthogonal' to predictability based on prediction errors. Often, these two approaches disagree. For example, a Ko stochastic sequence (which is a notion of 'unpredictability') may still be predictable in the Tadaki sense.

Proposition 122. *There is a sequence A which is Ko stochastic and Tadaki predictable.*

Proof. Consider a Ko stochastic sequence A and construct a sequence B as follows

$$B_i = \begin{cases} 0 & \text{if } i = 2^j \text{ for some } j \quad (7.1) \\ A_i & \text{otherwise.} \quad (7.2) \end{cases}$$

Observe that B is also Ko stochastic. On the other hand, we have a T-predictor f

$$f(\sigma) = \begin{cases} 0 & \text{if } |\sigma| \neq 2^j \text{ for any } j \quad (7.3) \\ \square & \text{otherwise.} \quad (7.4) \end{cases}$$

Note that f is always correct when it do not return $\square$. So, A is Tadaki predictable. □

On the one hand, we are tempted to say that the sequence for which we can have a prediction error approaching 0 is quite 'predictable'. But such sequence does not have to be Tadaki predictable.

Proposition 123. *There is a sequence A and a predictor f such that $\varsigma(f, A) = 0$ but A is not Tadaki predictable.*

Proof. We will construct A by direct diagonalization. Let $h_0, h_1, \ldots$ be a listing of all T-predictors. The construction proceeds by stages. We start with $\alpha_{(0)} = \emptyset$. At each stage $s + 1$ we construct a finite sequence $\alpha_{(s+1)} \succeq \alpha_{(s)}$. The constructed sequence is defined by $A = \bigcup_{s=0}^{\infty} \alpha_{(s)}$.

Construction.
Stage 0: $\alpha_{(0)} = \emptyset$.
Stage s+1: Check whether there exists $\tau = \alpha_{(s)}00\ldots0$ such that $h_s(\tau) \neq \square$ and $|\tau| - |\alpha_{(s)}| \geq 2^{s+1}$. If so, take the minimal such τ and let

$$\alpha_{(s+1)} = \begin{cases} \tau 0 & \text{if } h_s(\tau) = 1 \quad (7.5) \\ \tau 1 & \text{if } h_s(\tau) = 0. \quad (7.6) \end{cases}$$

Otherwise, $\alpha_{(s+1)} = \alpha_{(s)} \underbrace{00\ldots0}_{2^{s+1}}$.

Verification.
Now, observe that for every T-predictor h, if there exists such k that $h(A_1^k) \neq \square$ then there also exists such k that $h(A_1^k) \neq A_{k+1}$. Therefore, A is not Tadaki-predictable.
Furthermore, we know that for every step s of the construction $|\alpha_{(s+1)}| \geq (|\alpha_{(s)}| + 2^{s+1})$ and $\#_1(\alpha_{(s+1)}) \leq (\#_1(\alpha_{(s)}) + 1)$. As a consequence, $\lim_{n\to\infty} \#_1(A) = 0$. Consider a predictor χ such that $\chi[2^{<\mathbb{N}}] = \{0\}$ and observe that $\varsigma(\chi, A) = 0$. □

We can observe that every c.e. sequence is Tadaki predictable. Therefore, there is a Tadaki predictable sequence that is not error stable.

Proposition 124. *If A is c.e. then A is Tadaki predictable.*

Proof. Let A be a c.e. sequence. Since A is c.e., there exists an infinite set $B \subset A$ such that B is computable. Therefore, we have a T-predictor f

$$f(\sigma) = \begin{cases} 1 & \text{if } B_{|\sigma|+1} = 1 \quad (7.7) \\ \square & \text{otherwise.} \quad (7.8) \end{cases}$$

Observe that f is always correct when it places a nonzero bit after some prefix. Thus, A is Tadaki predictable. □

Corollary 125. *There is a Tadaki predictable sequence which is not error stable.*

Proof. By Theorem 106, there is a c.e. sequence that is not error stable. By Proposition 124, every c.e. sequence is Tadaki predictable. □

In fact, the Proposition 124 can be easily generalized to n-c.e. case. On the other hand, we know that a Church stochastic sequence is not Tadaki predictable. Recall that there is a Church stochastic sequence in ω-c.e. degrees. So

Corollary 126. *There is an ω-c.e. sequence that is not Tadaki predictable.*

We have already managed to show that Church stochasticity implies unpredictability in the Tadaki sense. Therefore, we know that Church stochasticity implies both Ko stochasticity and Tadaki unpredictability. It remains to check if the converse implication is true. Unfortunately, it is not.

Proposition 127. *There exists a sequence A that is Ko stochastic and not Tadaki predictable but is not Church stochastic.*

Proof. We will construct a Ko stochastic sequence X. On most of the bits we will just use the standard Ville's construction (see the proof of the Theorem 67). However, on every bit with index of form 2^i for some $i \in \mathbb{N}$ instead of going through Ville's construction, we will chose bits according to the procedure that we have used in the proof of Proposition 123.
Strictly speaking, we will construct a sequence A. Let $h_0, h_1, \ldots$ be a listing of all T-predictors. The construction proceeds by stages. We start with $\alpha_{(0)} = \emptyset$. At each stage $s+1$ we construct a finite sequence $\alpha_{(s+1)} \succeq \alpha_{(s)}$. The constructed sequence is defined by $A = \bigcup_{s=0}^{\infty} \alpha_{(s)}$. We will also use auxiliary variables e (which will be used to denote the index of a T-predictor) and c (a counter).

Construction.
Stage 0: $\alpha_{(0)} = \emptyset$, $e = 0$, $c = 0$.
Stage s+1: Check if $s = 2^i$ for some $i \in \mathbb{N}$. If not, simply act as in the Ville construction. Otherwise, check if $h_e(\alpha_s) = \square$. If so, let $\alpha_{(s+1)} = \alpha_{(s)}0$ and increase c by 1 (i.e. $c := c + 1$).
Otherwise, check if $c > 1$. If so, let

$$\alpha_{(s+1)} = \begin{cases} \alpha_{(s)}0 & \text{if } h_s(\alpha_{(s)}) = 1 \quad (7.9) \\ \alpha_{(s)}1 & \text{if } h_s(\alpha_{(s)}) = 0. \quad (7.10) \end{cases}$$

Additionally, reset the counter c to 0 and increase e by 1.
In case that $c > 1$ did not hold, let $\alpha_{(s+1)} = \alpha_{(s)}0$.

Verification.
Observe that A is infinite, since at every step of construction $|\alpha_{(s+1)}| > |\alpha_{(s)}|$. Moreover, A was constructed in such a way that A is stochastic in respect to every computable selection function f such that $f(\sigma) = 0$ if $|\sigma| = 2^i$ for some $i \in \mathbb{N}$. We already know that in such a case, A is Ko stochastic. (See Theorem 66 and Corollary 69.

Furthermore, consider three indexes i, j, k such that for some $m \in \mathbb{N}$ we have $i = 2^m$, $j = 2^{m+1}$ and $k = 2^{m+2}$. Note that at least two bits in bits of indexes i, j, k are zero (we used the counter c to ensure that — a nonzero bit could be placed only after c

was increased twice after the last reset or increased twice and never reset yet). Then, consider a selection function g such that $g(\sigma) = 1$ if and only if $|\sigma| = 2^i$ for some $i \in \mathbb{N}$ and observe that the subsequence selected by g on A is not balanced (between every nonzero bit there are at least two zero bits). So, A is not Church stochastic.

It remains to observe that for every T-predictor h there exists such k that $h(A_1^k) \neq \square$ but also $h(A_1^k) \neq A_{k+1}$. Therefore, A is not Tadaki-predictable. $\square$

7.2 Coarse computability

The notion of coarse computability was firstly introduced by Jokusch and Schupp in [21] and was subsequently studied in, e.g., [18][4][27] .

Definition 128 (Coarse computability)**.** *A sequence A is said to be coarsely computable in density r if there exists a computable $f : \mathbb{N} \to \mathbb{N}$ such that*

$$\liminf_{n\to\infty} \frac{\#_1\{n : f(n) = A_n\}}{n} = r.$$

Moreover, we will simply say that A is coarsely computable if it is coarsely computable in density 1*.*

Definition 129 (Coarse computability bound)**.** *A coarse computability bound $\gamma(A)$ of a sequence A is defined as the lowest upper bound for r such that A is coarsely computable in density r.*

If an uncomputable set A is coarsely computable then it means it can be well approximated with some computable set. In a way, this allows us to neglect uncomputability of A. If that is not the case, then γ measures how far A is from being coarsely computable.

Coarse computation may be recognized as a special case of prediction, i.e., such where the predictor trajectories only depend on the length of the input prefixes. Thus, we get a simple observation

Observation 130. *Let $A \in 2^{\mathbb{N}}$ and suppose that A is coarsely computable in density r. There exists a predictor f such that $\varsigma_-(f, A) = r$.*

But we also can easily get a separation result for these two notions.

Proposition 131. *There exists a sequence X such that $\gamma(X) = 1/2$ but X is predictable with error* $1/4$*.*

Proof. By Ville's theorem, for every $\mathcal{F}$ — a countable collection of functions from words into $\{0, 1\}$ we can construct a sequence A such that every f from $\mathcal{F}$ we have $\varsigma(f, A) = 1/2$. Let $\mathcal{F}$ be a collection of all computable functions f such that for each σ, τ

$$|\sigma| = |\tau| \Rightarrow f(\sigma) = f(\tau).$$

Let A be a corresponding sequence stochastic with respect to $\mathcal{F}$. Note that $\gamma(A) = 1/2$. Consider $X = A \oplus A$. We are going to prove that $\gamma(X) = 1/2$. Suppose otherwise. Then, there exists a computable function f which select an unbalanced sequence on A. But then f selects an unbalanced subsequence on all odd or all even indexes. Therefore, we can easily construct a function g which selects an unbalanced subsequence on A, a contradiction.
Subsequently, contemplate the function g such that

$$g(\sigma) = \begin{cases} 0 & \text{if } |\sigma| = 2k \quad (7.11) \\ \sigma_{|\sigma|} & \text{if } |\sigma| = 2k+1 \quad (7.12) \end{cases}$$

On even indexes, g behaves as a function from $\mathcal{F}$ — so it makes exactly $1/2$ correct answers on these bits, by the construction of A and X. On odd indexes g simply copies the last bit of the prefix and hence it is always correct there. Consequently, $\varsigma(f, X) = 1/4$. □

Problem 132. *Can we strengthen this to get $\varsigma(f, X) = 0$?*

An important question in coarse computability concerned the following concept.

Definition 133 (Big gamma)**.** *Let* $\mathbf{a}$ *be a Turing degree.*

$$\Gamma(\mathbf{a}) = \inf\{\gamma(A) : A \text{ is } \mathbf{a}\text{-computable}\}$$

.

Theorem 134 (Monin [27])**.** *There is no degree* $\mathbf{a}$ *such that* $0 < \Gamma(\mathbf{a}) < 1/2$.

We can define an analogue of Γ for prediction.

Definition 135 (Lower predictability bound)**.** *A predictability bound $\xi(A)$ of a sequence A is defined as the lowest upper bound for r such that there exists a predictor f with $\varsigma_-(f, A) = r$.*
Furthermore, for a degree $\mathbf{a}$ *let*

$$\Xi(\mathbf{a}) = \inf\{\xi(A) : A \text{ is } \mathbf{a}\text{-computable}\}$$

Question 136. *What can we say about Ξ? Does it behave similar to Γ?*

Part III

Prediction in probabilistic framework

8

Randomness and nonuniform measures

Such was of old, quoth Epistemon, the custom of the grand vaticinator and prophet Tiresias, who used always, by way of a preface, to say openly and plainly at the beginning of his divinations and predictions that what he was to tell would either come to pass or not. And such is truly the style of all prudently presaging prognosticators. He was nevertheless, quoth Panurge, so unfortunately misadventurous in the lot of his own destiny, that Juno thrust out both his eyes.

– François Rabelais, The Third Book[1]

8.1 Randomness and nonuniform measures

The familiar notion of Martin-Löf tests and Martin-Löf randomness may be generalized to arbitrary computable measures on $2^{\mathbb{N}}$ in a very natural way.

Definition 137 (uniformly Σ^0_1 collection of sets of sequences)**.** *A collection $U_0, U_1, \ldots$ of sets of sequences is uniformly Σ^0_1 if and only if there is a collection $V_0, V_1, \ldots \subset 2^{<\mathbb{N}}$ such that $U_i = [\![V_i]\!]$ for every $i \in \mathbb{N}$ and $V_0, V_1, \ldots$ are uniformly Σ^0_1.*

Definition 138 (Martin-Löf μ-test)**.** *Let μ be a computable measure on $2^{\mathbb{N}}$. A uniformly Σ^0_1 sequence $U_0, U_1, \ldots$ of sets of sequences is called a Martin-Löf μ-test if there exists a computable f such that $\lim_{n\to\infty} f(n) = 0$ and $\mu(U_n) \leq f(n)$ for every $n \in \mathbb{N}$.*

Definition 139 (Martin-Löf μ-randomness)**.** *Let μ be a computable measure on $2^{\mathbb{N}}$. A sequence $A \in 2^{\mathbb{N}}$ is called Martin-Löf μ-random (or simply, μ-random) if there is no such Martin-Löf μ-test $U_1, U_2 \ldots$ that $A \in \bigcap_{i\in\mathbb{N}} U_n$.*

So far, we have dealt only with one-sided infinite binary sequences. We can, however, think of these as subsequences of some two-sided infinite sequences. This will prove

[1] as translated by Thomas Urquhart and Peter Antony Motteux.

useful as we go further and turn our attention to stationary measures. It is known that every stationary measure on one-sided infinite sequences may be uniquely extended to a measure on two-sided infinite sequences. To make a proper use of this fact, we first need to extend the conceptual apparatus of algorithmic randomness to the universe of two-sided infinite sequences.

While warming-up, observe that every two-sided infinite binary sequence y corresponds to a one-sided infinite sequence x in a simple manner. For each $k \in \mathbb{N}$ let

$$y_k = x_{2k+1}$$

and for all $k \in \mathbb{N}^+$

$$y_{-k} = x_{2k}.$$

Hence, computability-theoretic notions are easily transfered onto two-sided infinite sequences. We introduce the following notation for cylinder sets of two-sided infinite sequences. For all $w \in 2^{<\mathbb{N}}$ with $|w| = 2k+1$ for some $k \geq 0$

$$[\![w]\!]_{\mathbb{Z}} = \{x \in 2^{\mathbb{Z}} : x_{-k}^{k} = w\}.$$

Similarly, for a set of words $S \subset 2^{<\mathbb{N}}$ let

$$[\![S]\!]_{\mathbb{Z}} = \bigcup_{w \in S} [\![w]\!]_{\mathbb{Z}}.$$

Definition 140 (uniformly Σ_1^0 collection of sets of two-sided sequences)**.** *A collection $U_0, U_1, \ldots$ of sets of two-sided infinite sequences is uniformly Σ_1^0 if and only if there is a collection $V_0, V_1, \ldots \subset 2^{<\mathbb{N}}$ such that $U_i = [\![V_i]\!]_{\mathbb{Z}}$ for every $i \in \mathbb{N}$ and $V_0, V_1, \ldots$ are uniformly Σ_1^0.*

Definition 141 (two-sided Martin-Löf μ-test)**.** *Let μ be a computable measure on $2^{\mathbb{Z}}$. An uniformly Σ_1^0 sequence $U_0, U_1, \ldots$ of sets of two-sided infinite sequences is called a Martin-Löf μ-test if there exists computable f such that $\lim_{i\to\infty} f(i) = 0$ and $\mu(U_i) \leq f(i)$ for every $i \in \mathbb{N}$.*

Definition 142 (two-sided Martin-Löf μ-randomness)**.** *Let μ be a computable measure on $2^{\mathbb{Z}}$. A two-sided infinite sequence $A \in 2^{\mathbb{Z}}$ is called Martin-Löf μ-random (or simply, μ-random) if there is no such two-sided infinite Martin-Löf μ-test $U_0, U_1 \ldots$ that $A \in \cap_{i\in\mathbb{N}} U_n$.*

8.2 Effective almost-everywhere theorems

Let μ be a probability measure on infinite sequences. In modern probability theory, many results are stated in the following form

$$\mu(\{\omega : \phi(\omega)\}) = 1,$$

where ϕ is some formula — often stating a pointwise convergence. The above is usually stated as '$\phi(\omega)$ for μ-almost every ω'. In the algorithmic approach to probability, we seek for an effective version of such theorems, that is we want to see if

$$\phi(\omega) \text{ for every } \mu\text{-random } \omega.$$

(Note that the set of all μ-random sequences is of measure one.) Such an effectivization may be done for various different notions of randomness. In the following chapters we will seek effective theorems formulated for Martin-Löf randomness. From historical perspective, this research program originates from von Mises's concept of *Kollektive*. For an overview see [41].

8.3 Effective Borel-Cantelli lemma

As in the case of randomness for the uniform measure, besides the definition by Martin-Löf tests, we have an equivalent definition based on Solovay tests. We note that this observation may be seen as an effective Borel-Cantelli lemma.

Definition 143 (Solovay test)**.** *Let μ be a computable measure. A Solovay μ-test is a collection $S_0, S_1, \ldots$ of uniformly Σ^0_1 sets of sequences such that $\sum_{n\in\mathbb{N}} \mu(S_n) < \infty$. A sequence $x \in 2^{\mathbb{Z}}$ is called Solovay μ-random if for every such test, x belongs only to finitely many S_n.*

Theorem 144 (Solovay, unpublished)**.** *A sequence is Martin-Löf μ-random if and only if it is Solovay μ-random.*

Proof. See e.g., Theorem 6.2.8 in [12]. □

8.4 Martingale convergence

As an example of the effectivization, we will now state an effective version of Doob's martingale convergence theorem. Typically, to obtain an effective theorem of form '$\phi(\omega)$ for all μ-random ω' we have to to put some computability constraint on ϕ. Hence, to get an effective martingale convergence theorem, we need to define a notion of a computable martingale process. The standard notion of a computable function applies to mappings from countable sets into countable sets. On the other hand, a martingale process is a function which acts on points of the probability space, i.e., on infinite sequences. A definition of a computable martingale process was proposed, for example, by Takahashi [37], who also proved a version of an effective Doob's theorem. Given a process X, Takahashi simply requires that for all numbers n, the value X_n depends only on the first n bits of the sequence. This definition is rather crude and not sufficient for our purpose. We want the process to depend on a finite fragment of a sequence but we do not want to bound the length of that fragment a priori. An elegant way to introduce computability with respect to a finite but potentially infinite input is given by oracle machines, as exemplified by the following definitions.

Definition 145 (sequences as oracles)**.** *Fix a computable bijection $h : \mathbb{N} \to \mathbb{Z}$. Choose $\omega \in 2^{\mathbb{Z}}$. We will say that a set is computable with oracle ω if and only if it is computable with oracle W, where*

$$W = \{n \in \mathbb{N} : \omega_{h(n)} = 1\}.$$

Definition 146 (computable martingale process)**.** *We will say that $X = X_1, X_2, \ldots$ is a computable martingale process if it is a martingale process and there exists a program for the register oracle machine which computes the set*

$$\{(n, q) : n \in \mathbb{N}^+ \wedge q \in \mathbb{Q} \wedge q < X_n(\omega)\}$$

for each oracle ω.

The martingale process may be seen as an algorithm acting on some fragment of the sequence ω to produce a rational approximation of a real value (the left cut). Before we can prove the effective Doob's theorem, we need to recall an important probabilistic lemma.

Lemma 147. *Let $X = X_1, X_2, \ldots$ be a martingale process and let C_n will be the random variable denoting the number of upcrossings of interval $[a, b]$ (with $a, b \in \mathbb{R}$) by time n, i.e., it is the largest number t such that*

$$1 \leq l_1 < u_1 < l_2 < u_2 < \ldots < l_t < u_t \leq n$$

where for each $i \leq t$ we have $X_{l_i} < a$ and $X_{u_i} > b$. Moreover, suppose that $\sup_n \mathbb{E}(|X_n|) < \infty$. Then the following holds for each n

$$\mathbb{E}(C_n) \leq \frac{|a| + \mathbb{E}(|X_n|)}{b - a}.$$

Proof. See e.g. [8]. □

Applying the monotone convergence theorem we get

Corollary 148. *Let $X = X_1, X_2, \ldots$ be a martingale process and let C_n will be the random variable denoting the number of upcrossings of interval $[a, b]$ (with $a, b \in \mathbb{R}$) by time n and suppose that $\sup_n \mathbb{E}(|X_n|) < \infty$. Then the following holds for each n*

$$\mathbb{E}(\sup_n C_n) \leq \frac{|a| + \sup_n \mathbb{E}(|X_n|)}{b - a}.$$

Theorem 149 (effective Doob's martingale convergence)**.** *Set a measure μ on two-sided infinite sequences and let $Y = Y_1, Y_2, \ldots$ be a computable martingale process. Then, the limit $\lim_{n\to\infty} Y_n$ exists for each μ-random ω.*

Proof. Fix arbitrary $a, b \in \mathbb{Q}$ and let C_n be the random variable denoting the number of upcrossings of interval $[a, b]$ by the process Y by the time n. Let C_∞ denote $\sup_n C_n$. Consider a collection of sets $U = U_0, U_1, \ldots$ such that for all $i \in \mathbb{N}^+$

$$U_i = \{\omega \in 2^{\mathbb{Z}} : C_\infty(\omega) > i\}$$

By Corollary 148 and the Markov inequality, we have

$$\mu(U_i) \leq \frac{|a| + \sup_n \mathbb{E}(|Y_n|)}{i(b-a)}$$

To see that U is a Martin-Löf μ-test we only need to argue that collection $U_0, U_1, \ldots$ is uniformly Σ_1^0. Since Y is a computable martingale process then given an $\omega \in 2^{\mathbb{Z}}$ and $k \in \mathbb{N}$, the question whether $C_k(\omega) > i$ is computable with oracle ω. Moreover, the answer depends only on a finite fragment of ω. It follows that we can effectively enumerate elements of the set V_i generating $U_i = [\![V_i]\!]$. Indeed, we can fix an enumeration of words and iterate via pairs $(w, k) \in 2^{<\mathbb{N}} \times \mathbb{N}$. At each step we ask whether $C_k(x) > i$, where $x = \ldots 00w00 \ldots$ (with $x_{-\lfloor |w|/2 \rfloor}^{\lceil |w|/2 \rceil} = w$). If so, then we include w into V_i.

Finally, since a, b are arbitrary and U is a Martin-Löf μ-test, it follows that $C_\infty(\omega) < \infty$ for every μ-random ω and hence, $\lim_{n\to\infty} Y_n$ exists for all μ-random sequences. □

Again, we have a version of Lévy's law as a corollary.

Theorem 150 (effective Lévy's law)**.** *Let X be a two-sided infinite process with the computable probability distribution μ. Let $\mathcal{F}_0, \mathcal{F}_1, \ldots$ be a filtration and define $\mathcal{F}_\infty$ to be σ-algebra generated by the union $\bigcup_{n\in\mathbb{N}} \mathcal{F}_n$. For every μ-random sequence $x \in 2^{\mathbb{Z}}$ and for each $k \in \mathbb{N}$, $\lim_{n\to\infty} \mathbb{E}(X_k|\mathcal{F}_n)$ exists and is a version of $\mathbb{E}(X_k|\mathcal{F}_\infty)$.*

8.5 Stationary ergodic processes

In this section a certain class of processes is introduced, namely, the class of stationary ergodic processes. These processes have received much attention in probability and information theory. This attention comes, in particular, from the group of results called the ergodic theorems (see Section 8.6). These theorems guarantee that empirical realizations of the processes have well-behaved statistical properties. As we will see later on, we can use it as an advantage for prediction of such processes.

Definition 151 (measure preserving transformation)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space. A measurable transformation $T : \Omega \to \Omega$ is said to preserve measure P if and only if for all $A \in \mathcal{F}$*

$$P(T^{-1}(A)) = P(A).$$

Definition 152 (stationary measure and process)**.** *Let X be a one-sided (two-sided) infinite binary process with a probability distribution μ. Let $I = \mathbb{N}$ or $I = \mathbb{Z}$. Let T be a right shift, that is for all $k \in \mathbb{N}^+$ ($k \in \mathbb{Z}$) and every $\omega \in 2^{\mathbb{N}}$ ($\omega \in 2^{\mathbb{Z}}$):*

$$T(\omega)_k = \omega_{k+1}.$$

We will say that μ is stationary if T preserves μ. We will also say that X is a stationary process if its probability distribution is stationary.

Definition 153 (ergodic transformation)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space. A measure preserving transformation $T : \Omega \to \Omega$ is called ergodic if and only if for each $A \in \mathcal{F}$ such that $T^{-1}(A) = A$ either $P(A) = 1$ or $P(A) = 0$.*

Definition 154 (ergodic measure and process)**.** *Let X be a one-sided (two-sided infinite) binary process with a probability distribution μ. We will say that μ is ergodic if the right shift is an ergodic transformation with respect to μ. Similarly, the process is called ergodic if its probability distribution is ergodic.*

Any stationary measure defined for one-sided infinite sequences may be uniquely extended to a stationary measure on two-sided infinite sequences.

Proposition 155 (Kolmogorov process theorem II)**.** *If ν is a stationary probability measure on one-sided infinite sequences, then there exists a unique process X with probability distribution μ on two-sided infinite sequences such that for every $i \in \mathbb{Z}$ and $x_1^n \in 2^{<\mathbb{N}}$*

$$\mu(X_{i+1}^{i+n} = x_1^n) = \nu(x_1^n)$$

The following proposition allow us to transfer various results about algorithmically random two-sided infinite sequences to the realm of one-side infinite sequences.

Proposition 156. *Let X be a two-sided infinite stationary process with probability distribution μ and let ν be a stationary measure on one-sided infinite sequences such that for every $i \in \mathbb{N}$ and $x_1^n \in 2^{<\mathbb{N}}$*

$$\mu(X_{i+1}^{i+n} = x_1^n) = \nu(x_1^n).$$

Suppose that $x \in 2^{\mathbb{Z}}$ is μ-random. Then $y = x_1^\infty$ is ν-random.

Proof. Fix measures μ and ν on two-sided and one-sided infinite sequences respectively. Suppose that $x \in 2^{\mathbb{Z}}$ is μ-random but x_1^∞ is not ν-random. Hence, there exists $[\![V_1]\!], [\![V_2]\!], \ldots$ — a ν-test which witnesses non-randomness of x_1^∞. Now, suppose that μ and ν satisfy

$$\mu(X_{i+1}^{i+n} = y_1^n) = \nu(y_1^n)$$

for every $i \in \mathbb{N}$ and $y_1^n \in 2^{<\mathbb{N}}$. Let define U_i for each $i \in \mathbb{N}$ as a set of all two-sided infinite sequences satisfying $X_1^\infty = y$ for all $y \in [\![V_i]\!]$. In other words, we take all the generators for V_i (which is the set of one-sided infinite sequences) and use them to obtain cylinder sets on two-sided infinite sequences. The collection $U_1, U_2, \ldots$ forms a μ-test which witnesses that x is not μ-random. □

8.6 Ergodic theorems

The fundamental ergodic theorem is Birkhoff's ergodic theorem

Theorem 157 (Birkhoff's ergodic theorem [9])**.** *Let μ be a stationary ergodic measure on binary sequences and let $f : 2^{\mathbb{Z}} \to \mathbb{R}$ be measurable and integrable. Then for almost every sequence $\omega \in 2^{\mathbb{Z}}$*

$$\lim_{i \to \infty} \frac{1}{i} \sum_{k=1}^{i} f(T^k(\omega)) = \mathbb{E}(f).$$

The above theorem has an effective version

Theorem 158 (effective Birkhoff's ergodic theorem [15][7])**.** *Let μ be a computable stationary ergodic measure on binary sequences and let $f : 2^{\mathbb{Z}} \to \mathbb{R}^+$ be left c.e.. For every μ-random sequence $\omega \in 2^{\mathbb{Z}}$,*

$$\lim_{i\to\infty} \frac{1}{i} \sum_{k=1}^{i} f(T^k(\omega)) = \mathbb{E}(f).$$

Breiman [10] proved a generalization of Birkhoff's ergodic theorem. We start by stating its classical version.

Theorem 159 (Breiman's ergodic theorem [10])**.** *Let μ be a stationary ergodic measure on binary sequences and let $g_1, g_2, \ldots : 2^{\mathbb{Z}} \to \mathbb{R}$ such that the limit $\lim_{i\to\infty} g_i$ exists almost surely and $\mathbb{E}(\sup_i |g_i|) < \infty$. For almost every sequence $\omega \in 2^{\mathbb{Z}}$,*

$$\lim_{i\to\infty} \frac{1}{i} \sum_{k=1}^{i} g_k(T^k(\omega)) = \mathbb{E}(\lim_{k\to\infty} g_k).$$

Similarly, we can derive an effective version of Breiman's theorem

Theorem 160 (effective Breiman's ergodic theorem)**.** *Let μ be a computable stationary ergodic measure on binary sequences and let $g_1, g_2, \ldots : 2^{\mathbb{Z}} \to \mathbb{R}^+$ such that g_i are uniformly computable and $\mathbb{E}(\sup_i |g_i|) < \infty$. Then for every μ-random sequence $\omega \in 2^{\mathbb{Z}}$,*

$$\limsup_{i\to\infty} \frac{1}{i} \sum_{k=1}^{i} g_k(T^k(\omega)) \leq \mathbb{E}(\limsup_{k\to\infty} g_k)$$

and

$$\liminf_{i\to\infty} \frac{1}{i} \sum_{k=1}^{i} g_k(T^k(\omega)) \geq \mathbb{E}(\liminf_{k\to\infty} g_k).$$

Proof. Let $G_k = \sup_{t>k} g_t$. If so, then $g_t \leq G_k$ for all $t > k$ and consequently,

$$\limsup_{i\to\infty} \frac{1}{i} \sum_{k=1}^{i} g_k(T^k(\omega)) \leq \lim_{i\to\infty} \frac{1}{i} \sum_{k=1}^{i} G_k(T^k(\omega)).$$

Observe that a supremum of uniformly computable functions $g_k, g_{k+1}, \ldots$ is left-c.e.. Indeed, to enumerate the left cut of the supremum we simply simultaneously enumerate left cuts of $g_k(\omega), g_{k+1}(\omega), \ldots$. This is possible since every computable function is also c.e.. Moreover, we are considering only countably many functions, hence we can guarantee that an element of each left cut appears in the enumeration at least once.

Now, since for all $j \in \mathbb{N}$, G_j is left-c.e., by the effective Birkhoff's ergodic theorem, for every μ-random sequence ω,

$$\limsup_{i\to\infty} \frac{1}{i} \sum_{k=1}^{i} g_k(T^k(\omega)) \leq \mathbb{E}(G_j).$$

Random variables $G_1, G_2, \ldots$ are bounded from above by an integrable function. Hence, by the dominated convergence theorem,

$$\lim_{i\to\infty} \mathbb{E}(G_i) = \mathbb{E}(\lim_{i\to\infty} G_i) = \mathbb{E}(\limsup_{t\to\infty} g_t).$$

Thus,

$$\limsup_{i\to\infty} \frac{1}{i} \sum_{k=1}^{i} g_k(T^k(\omega)) \leq \mathbb{E}(\limsup_{t\to\infty} g_t).$$

For the second inequality, consider variables $H_k = \sup_{t>k} -g_t$. Following similar reasoning as above, we obtain for every μ-random ω

$$\limsup_{i\to\infty} \frac{1}{i} \sum_{k=1}^{i} -g_k(T^k(\omega)) \leq \mathbb{E}(\limsup_{t\to\infty} -g_t).$$

This is equivalent to

$$\liminf_{i\to\infty} \frac{1}{i} \sum_{k=1}^{i} g_k(T^k(\omega)) \geq \mathbb{E}(\liminf_{k\to\infty} g_k).$$

□

9

Lower bounds for zero-one loss

Suppose that the measure of the underlying process is known in advance. It seems that the optimal strategy should simply consist in predicting the more probable outcome. Indeed, in what follows we will prove that this is the case. While, in general, the measure is usually not known, it does provide a lower bound on the prediction error for every proper predictor.

Lemma 161. *Let $X = X_1, \dots$ be a binary process with a probability distribution μ. The following is satisfied:*

$$\lim_{n\to\infty} \left| \varsigma(f, x_1^n) - \frac{1}{n} \sum_{i=1}^{n} \mu(f(X_1^{i-1}) \neq X_i | X_1^{i-1} = x_1^{i-1}) \right| = 0$$

for each μ-random x and computable f.

Proof. Fix X, μ, g and f as needed. Consider a process $Y = Y_0, Y_1, \dots$ with $Y_0 = 0$ and

$$Y_n = n\varsigma(f, X_1^n) - \sum_{i=1}^{n} \mu(f(X_1^{i-1}) \neq X_i | X_1^{i-1}).$$

Note that

$$\mathbb{E}(Y_{n+1} | X_1^n) = Y_n + \mathbb{E}\left(|f(X_1^n) - X_{n+1}| - \mu(f(X_1^n) \neq X_{n+1} | X_1^n) | X_1^n \right) = Y_n.$$

Hence, Y is a martingale process. Fix an $\epsilon > 0$. By Azuma's inequality, for $t > 0$, we obtain

$$\mu\left(|Y_n - Y_0| > t \right) \leq 2e^{-2t^2/n}.$$

Consequently, for each $t > 0$

$$\mu\left(|Y_n - Y_0| > nt \right) \leq 2e^{-2nt^2}.$$

Fix $\epsilon > 0$ and set $t = n\epsilon$ to obtain

$$\mu\left(\left| \varsigma(f, X_1^n) - \frac{1}{n} \sum_{i=1}^{n} \mu(f(X_1^{i-1}) \neq X_i | X_1^{i-1}) \right| > \epsilon \right) \leq 2e^{-2n\epsilon^2}.$$

It remains to consider the following family of sets $U = [\![U_0]\!], [\![U_1]\!], \ldots$

$$U_n = \left\{ x \in 2^{<\mathbb{N}} : \left| \varsigma(f, x_1^n) - \frac{1}{n} \sum_{i=1}^{n} \mu(f(X_1^{i-1}) \neq X_i | X_1^{i-1} = x_1^{i-1})) \right| > \epsilon \right\}.$$

Since $\mu(U_n)$ is bounded by a summable function, U forms a Solovay μ-test. Suppose that for some $x \in 2^{\mathbb{Z}}$

$$\limsup_{n\to\infty} \left| \varsigma(f, x_1^n) - \frac{1}{n} \sum_{i=1}^{n} \mu(f(X_1^{i-1}) \neq X_i | X_1^{i-1} = x_1^{i-1}) \right| > 0.$$

But then, if ϵ is sufficiently small, then the sequence x is in infinitely many $[\![U_i]\!]$. Since U is a Solovay test, we conclude that x is not μ-random. This ends the proof. □

Theorem 162. *Let X be a binary process with a probability distribution μ. Let predictor h be defined as*

$$h(X_1^n) = 1 \Leftrightarrow \mu(X_{n+1} = 1 | X_1^n) > 1/2.$$

Then for every proper predictor f and every μ-random $x \in 2^{\mathbb{Z}}$

$$\liminf_{n\to\infty} (\varsigma(f, x_1^n) - \varsigma(h, x_1^n)) \geq 0.$$

Proof. Fix a binary process X with the probability distribution μ and h defined as in the assumption of the theorem. Firstly, fix $x \in 2^{\mathbb{Z}}$. Obviously, for all $n \in \mathbb{N}^+$ either $f(x_1^n) = 0$ or $f(x_1^n) = 1$. Consequently, one of the following is true:

$$\mu(f(X_1^n) \neq X_{n+1} | X_1^n = x_1^n) = \mu(X_{n+1} = 1 | X_1^n = x_1^n)$$

or

$$\mu(f(X_1^n) \neq X_{n+1} | X_1^n = x_1^n) = \mu(X_{n+1} = 0 | X_1^n = x_1^n).$$

In other words, since h always chooses a more probable outcome, we have

$$\begin{aligned} \mu(f(X_1^n) \neq X_{n+1} | X_1^n = x_1^n) \geq & \mu(h(X_1^n) \neq X_{n+1} | X_1^n = x_1^n) \\ & = \min\{\mu(X_{n+1} = 1 | X_1^n = x_1^n), 1 - \mu(X_{n+1} = 1 | X_1^n = x_1^n)\} \end{aligned}$$

Consequently,

$$\frac{1}{n} \sum_{i=0}^{n-1} \mu(f(X_1^i) \neq X_{i+1} | X_1^i = x_1^i) \geq \frac{1}{n} \sum_{i=0}^{n-1} \mu(h(X_1^i) \neq X_{i+1} | X_1^i = x_1^i).$$

and so

$$\liminf_{n\to\infty} \left(\frac{1}{n} \sum_{i=0}^{n-1} \mu(f(X_1^i) \neq X_{i+1} | X_1^i = x_1^i) - \frac{1}{n} \sum_{i=0}^{n-1} \mu(h(X_1^i) \neq X_{i+1} | X_1^i = x_1^i) \right) \geq 0.$$

The claim follows by Lemma 161. □

10

Universal prediction

In the previous section we have seen that the best strategy is to base the prediction on the knowledge of the underlying measure. Such strategy is optimal but not necessarily computable. Moreover, in real scenarios, the underlying probability measure is usually unknown. Therefore, it is natural to ask about some form of a universal strategy that could guess the underlying measure from the experience. Is goes without saying, that there is no universal learning scheme for an arbitrary measure. Hence, we need to limit our attention to some class of processes, e.g. the class of stationary ergodic processes. In the following section we are going to describe an universal strategy and prove its optimality with respect to every stationary ergodic process.

10.1 Backward measure estimation

The following theorem was proven by Ornstein [32].

Theorem 163 (Ornstein [32])**.** *There exists a sequence of functions $g_1, g_2, \ldots$ such that for every two-sided infinite process $X = \ldots X_{-1}, X_0, X_1 \ldots$ with the stationary probability distribution μ,*

$$\lim_{n\to\infty} g_n(\omega_{-n}^{-1}) = \mu(X_0 = 1 | X_{-\infty}^{-1} = \omega_{-\infty}^{-1})$$

for μ-almost every sequence ω.

Our proof of the effective version of the theorem uses a simplified algorithm of Morvai-Yakovitz-Györfi [28].

Theorem 164. *There exists a sequence of functions $g_1, g_2, \ldots$ such that for every two-sided infinite process $X = \ldots X_{-1}, X_0, X_1 \ldots$ with the stationary probability distribution μ,*

$$\lim_{n\to\infty} g_n(\omega_{-n}^{-1}) = \lim_{n\to\infty} \mu(X_0 = 1 | X_{-n}^{-1} = \omega_{-n}^{-1})$$

for every μ-random sequence ω.

Proof. Firstly, we define sequences of random variables λ_{k-1} and τ_k (for $k = 1, 2 \ldots$).

Let $\lambda_0 = 1$ and for $k \geq 1$ let

$$\tau_k = \min\{t \geq 1 : X_{-\lambda_{k-1}-t}^{-1-t} = X_{-\lambda_{k-1}}^{-1}\}. \tag{10.1}$$

That is, τ_k is the time between the occurrence of word $X_{-\lambda_{k-1}}^{-1}$ at time -1 and the last occurrence of this word prior to time -1. Finally, let $\lambda_k = \tau_k + \lambda_{k-1}$. The sequence of functions $f_1, f_2, \ldots$ is then defined as

$$f_k(X_{-\tau_k}^{-1}) = \frac{1}{k} \sum_{1 \leq j \leq k} X_{-\tau_j}. \tag{10.2}$$

Now, we proceed to show that the sequence $f_1, f_2, \ldots$ is as needed. Consider the filtration $\mathcal{F}_0, \mathcal{F}_1, \ldots$ with $\mathcal{F}_j$ being the σ-algebra generated by sets of form

$$\{x \in 2^{\mathbb{Z}} : x_{-m}^{-1} = w \wedge \lambda_j = m\}$$

(with $w \in 2^{<\mathbb{N}}$ and $m \in \mathbb{N}$). For each $k \in \mathbb{N}$ we have

$$f_k(X_{-\tau_k}^{-1}) - \mu(X_0 = 1|X_{-\infty}^{-1}) = \frac{1}{k} \sum_{1 \leq j \leq k} X_{-\tau_j} - \mu(X_0 = 1|X_{-\infty}^{-1}) \tag{10.3}$$

$$= \frac{1}{k} \sum_{1 \leq j \leq k} (X_{-\tau_j} - \mu(X_{-\tau_j} = 1|\mathcal{F}_j) \tag{10.4}$$

$$+ \mu(X_{-\tau_j} = 1|\mathcal{F}_j) - \mu(X_0 = 1|X_{-\infty}^{-1})) \tag{10.5}$$

We start by considering a process Y with $Y_0 = 0$ and

$$Y_n = \sum_{1 \leq j \leq n} \left(X_{-\tau_j} - \mu(X_{-\tau_j} = 1|\mathcal{F}_j)\right) = \sum_{1 \leq j \leq n} \left(X_{-\tau_j} - \mathbb{E}(X_{-\tau_j}|\mathcal{F}_j)\right).$$

Note that

$$\mathbb{E}(Y_{n+1}|\mathcal{F}_n) = Y_n + \mathbb{E}(X_{-\tau_{n+1}} - \mu(X_{-\tau_{n+1}} = 1|\mathcal{F}_n)) = Y_n.$$

Hence, Y_n forms a martingale process with respect to $\mathcal{F}_0, \mathcal{F}_1, \ldots$. Fix an $\epsilon > 0$. Consider a collection of sets $U = [\![U_0]\!], [\![U_1]\!], \ldots$

$$U_i = \left\{ x \in 2^{\mathbb{N}} : \left| \frac{1}{k} \sum_{1 \leq j \leq k} \left(X_{-\tau_j} - \mu(X_{-\tau_j} = 1|\mathcal{F}_j)\right) \right| > \epsilon \right\}.$$

Applying Azuma's inequality we may conclude that U is a Solovay μ-test, so that

$$\lim_{k \to \infty} \frac{1}{k} \sum_{1 \leq j \leq k} \left(X_{-\tau_j} - \mu(X_{-\tau_j} = 1|\mathcal{F}_j)\right) = 0$$

is satisfied by all μ-random sequences. We continue by showing that for each natural k

$$\mu(X_{-\tau_j} = 1|\mathcal{F}_j) = \mu(X_0 = 1|\mathcal{F}_j).$$

Let T denote the right shift operator. Consider a sequence of random variables

$$\bar{\tau}_j = \min\{t > 0 : X^{-1+t}_{-\lambda_{j-1}+t} = X^{-1}_{-\lambda_{j-1}}\}$$

Observe that τ_i and $\bar{\tau}_i$ are defined in an analogous way. Indeed, while the former one is a return time measured to the left side of the sequence, the latter one is a return time measured to the right. Hence, in particular:

$$T^l(X^{-1}_{-m} = \sigma \wedge \lambda_{j-1} = m \wedge \tau_j = l \wedge X_{-l} = 1) = (X^{-1}_{-m} = \sigma \wedge \lambda_{j-1} = m \wedge \bar{\tau}_j = l \wedge X_0 = 1)$$

for all $j, m \in \mathbb{N}$ and each $\sigma \in 2^{<\mathbb{N}}$. Now, by stationarity of μ we may further observe that for all natural j, m and each $\sigma \in 2^{<\mathbb{N}}$

$$\begin{aligned}
&\mu(X^{-1}_{-m} = \sigma \wedge \lambda_{j-1} = m \wedge X_{-\tau_j} = 1) \\
&= \sum_{l=1}^{\infty} \mu(X^{-1}_{-m} = \sigma \wedge \lambda_{j-1} = m \wedge \tau_j = l \wedge X_0 = 1) \\
&= \sum_{l=1}^{\infty} \mu(T^l[X^{-1}_{-m} = \sigma \wedge \lambda_{j-1} = m \wedge \tau_j = l \wedge X_0 = 1]) \\
&= \sum_{l=1}^{\infty} \mu(X^{-1}_{-m} = \sigma \wedge \lambda_{j-1} = m \wedge \bar{\tau}_j = l \wedge X_0 = 1) \\
&= \mu(X^{-1}_{-m} = \sigma \wedge \lambda_{j-1} = m \wedge X_0 = 1)
\end{aligned}$$

Hence, for each natural k

$$\begin{aligned}
\mu(X_{-\tau_j} = 1|\mathcal{F}_j)(\omega) &= \mu(X_{-\tau_j} = 1|X^{-1}_{-\tau_j} = \omega^{-1}_{-\tau_j}) \\
&= \mu(X_0 = 1|X^{-1}_{-\tau_j} = \omega^{-1}_{-\tau_j}) \\
&= \mu(X_0 = 1|\mathcal{F}_j)(\omega).
\end{aligned}$$

Now, given a μ-random sequence, the values τ_j and λ_j are defined and finite for all j, by the effective version of Birkhoff's ergodic theorem. Note that if a sequence converges, then every its infinite subsequence converges too. Hence, by the effective version of Lévy's law,

$$\lim_{j\to\infty} \mu(X_0 = 1|\mathcal{F}_j)(\omega) = \lim_{n\to\infty} \mu(X_0 = 1|X^{-1}_{-n} = \omega^{-1}_{-n})$$

Now, functions $f_1, f_2, \ldots$ act on fragments $X^{-1}_{-\tau_1}, X^{-1}_{-\tau_2}, \ldots$ respectively. These may be straightforwardly used to construct a sequence of functions $g_1, g_2, \ldots$ acting on $X_{-1}, X^{-1}_{-2}, \ldots$ and so on. Simply, for each $j > 0$ and $\tau_j \leq k < \tau_{j+1}$ let $g(X^{-1}_{-k}) = f_j(X^{-1}_{-\tau_j})$. □

Bailey [6] showed that Ornstein's estimators may be also used in a forward fashion.

Theorem 165 (Bailey [6])**.** *There exists a sequence of functions $f_1, f_2, \ldots$ such that for every two-sided infinite process $X = \ldots X_{-1}, X_0, X_1 \ldots$ with the stationary ergodic distribution μ*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left| f_i(\omega_1^{i-1}) - \mu(X_i = 1|X_1^{i-1} = \omega_1^{i-1}) \right| = 0$$

for μ-almost every ω.

Theorem 166 (effective Bailey's theorem)**.** *There exists a sequence of functions $f_1, f_2, \ldots$ such that for every two-sided infinite process $X = \ldots X_{-1}, X_0, X_1 \ldots$ with the stationary ergodic distribution μ*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left| f_i(\omega_1^{i-1}) - \mu(X_i = 1 | X_1^{i-1} = \omega_1^{i-1}) \right| = 0$$

for each μ-random sequence ω.

Proof. By Theorem 164, there exists a sequence of functions $f_1, f_2, \ldots$ such that for every two-sided infinite process $X = \ldots X_{-1}, X_0, X_1 \ldots$ with corresponding computable probability measure μ

$$\lim_{n\to\infty} f_n(\omega_{-n}^{-1}) = \lim_{n\to\infty} \mu(X_0 = 1 | X_{-n}^{-1} = \omega_{-n}^{-1})$$

for each μ-almost every ω. Consequently,

$$\lim_{n\to\infty} |f_n(\omega_{-n}^{-1}) - \mu(X_0 = 1 | X_{-n}^{-1} = \omega_{-n}^{-1})| = 0$$

Note that the sequence of the absolute values in the above equation is bounded from above. Moreover, these are uniformly computable. Hence, we can apply the effective Breiman's ergodic theorem (Theorem 160) to get

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left| f_i(\omega_1^{i-1}) - \mu(X_i = 1 | X_1^{i-1} = \omega_1^{i-1}) \right| = \mathbb{E}(0) = 0.$$

for each μ-random ω. The claim follows from this immediately. □

10.2 A universal predictor

Combining earlier results we arrive at the finale. The following theorem follows from the results by Algoet [2].

Theorem 167. *There exists a predictor f such that for every stationary ergodic process $X = \ldots, X_{-1}, X_0, X_1, \ldots$, with a computable probability distribution μ, predictor f is optimal μ-almost surely.*

We now turn to prove the effective version of the Theorem 167.

Theorem 168. *There exists a proper predictor f such that for every stationary ergodic process $X = \ldots, X_{-1}, X_0, X_1, \ldots$, with the probability distribution μ, the predictor f is optimal for every μ-random sequence ω.*

Proof. We start by noting that for every predictor f

$$\begin{aligned} &\mu(f(X_1^n) \neq X_{n+1} | X_1^n) = 1 - \mu(f(X_1^n) = X_{n+1} | X_1^n) \\ &\leq 1 - \left(\mathbf{1}\{f(X_1^n) = 1\} \mu(X_{n+1} = 1 | X_1^n) + \mathbf{1}\{f(X_1^n) = 0\} (1 - \mu(X_{n+1} = 1 | X_1^n)) \right). \end{aligned} \tag{10.6}$$

Let h be defined as

$$h(X_1^n) = 1 \Leftrightarrow \mu(X_{n+1} = 1|X_1^n) > 1/2. \tag{10.7}$$

These yield, for arbitrary predictor f

$$\begin{aligned} &|\mu(f(X_1^n) \neq X_{n+1}|X_1^n) - \mu(h(X_1^n) \neq X_{n+1}|X_1^n)| \\ &\leq |\mu(X_{n+1} = 1|X_1^n)(\mathbf{1}\{h(X_1^n) = 1\} - \mathbf{1}\{f(X_1^n) = 1\}) \\ &+(1 - \mu(X_{n+1} = 1|X_1^n))(\mathbf{1}\{h(X_1^n) = 0\} - \mathbf{1}\{f(X_1^n) = 0\})| \\ &= 2|(\mu(X_{n+1} = 1|X_1^n) - 1)(\mathbf{1}\{h(X_1^n) = 1\} - \mathbf{1}\{f(X_1^n) = 1\})| \\ &\leq 2|\mu(X_{n+1} = 1|X_1^n) - 1/2|. \end{aligned} \tag{10.8}$$

Now, let $g_1, g_2, \ldots$ be the sequence of functions from the Theorem 164 and define a predictor f as

$$f(X_1^n) = 1 \Leftrightarrow g_n(X_1^n) > 1/2.$$

Note that $f(X_1^n) \neq h(X_1^n)$ implies

$$|\mu(X_{n+1} = 1|X_1^n) - g_n(X_1^n)| \geq |\mu(X_{n+1} = 1|X_1^n) - 1/2|. \tag{10.9}$$

Finally, fix $x \in 2^{\mathbb{Z}}$ and consider

$$\begin{aligned} &|\varsigma(f, x_1^{n+1}) - \frac{1}{n+1}\sum_{i=0}^{n} \mu(h(X_1^i) \neq X_{i+1}|X_1^i = x_1^i)| \\ &\leq |\varsigma(f, x_1^{n+1}) - \frac{1}{n+1}\sum_{i=0}^{n} \mu(f(X_1^i) \neq X_{i+1}|X_1^i = x_1^i)| \\ &+\frac{1}{n+1}\sum_{i=0}^{n} |\mu(f(X_1^i) \neq X_{i+1}|X_1^i = x_1^i) - \mu(h(X_1^i) \neq X_{i+1}|X_1^i = x_1^i)| \\ &\leq |\varsigma(f, x_1^{n+1}) - \frac{1}{n+1}\sum_{i=0}^{n} \mu(f(X_1^i) \neq X_{i+1}|X_1^i = x_1^i)| \\ &+\frac{2}{n+1}\sum_{i=0}^{n} |\mu(X_{i+1} = 1|X_1^i = x_1^i) - g_i(x_1^i)|. \end{aligned}$$

Now, the first term of the last upper bound converges to zero for each μ-random sequence x by Lemma 161. The second one has limit zero for each μ-random sequence by Theorem 166. The optimality of f follows from Lemma 162.

It remains to show that f is proper. In general, computability of g_n does not guarantee that f is computable (See Proposition 36). However, in our case f is indeed computable. Consult equations (10.1) and (10.2) to see that for each k the value of g_k is an average of finite number of integers. Hence, that number of bits needed to write the value of g_k down is bounded from above. Consequently, it is possible to check in finite number of steps if $g_k(\sigma) \geq 1/2$. Therefore, f is a proper predictor. □

11

Optimality

In this section, a probabilistic version of the results from Chapter 6 is presented.

Theorem 169. *There exists a binary process $X = X_1X_2...$ such that for every proper predictor f*

$$\varsigma_-(f, X) > 0 \text{ almost surely.} \tag{11.1}$$

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

$$\varsigma_+(f, X) < \epsilon \text{ almost surely.} \tag{11.2}$$

Consequently, for every proper predictor f there exists a proper predictor g such that

$$\varsigma_+(g, X) < \varsigma_-(f, X) \text{ almost surely.} \tag{11.3}$$

Proof. 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 112, 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 6.1 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

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

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

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

To demonstrate the second part of Theorem 169, 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,

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

□

Bibliography

[1] Ackerman, N.L., Freer, C.E., Roy, D.M.: On the computability of conditional probability. Journal of the ACM (JACM) 66(3), 23 (2019)

[2] Algoet, P.H.: The strong law of large numbers for sequential decisions under uncertainty. IEEE Transactions on Information Theory 40(3), 609–633 (1994)

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

[4] Astor, E.P., Hirschfeldt, D.R., Jockusch Jr, C.G.: Dense computability, upper cones, and minimal pairs. Computability (Preprint), 1–23 (2019)

[5] Azuma, K.: Weighted sums of certain dependent random variables. Tohoku Mathematical Journal, Second Series 19(3), 357–367 (1967)

[6] Bailey, D.H.: Sequential schemes for classifying and predicting ergodic processes. Ph. D Thesis, Stanford University (1976)

[7] Bienvenu, L., Day, A.R., Hoyrup, M., Mezhirov, I., Shen, A.: A constructive version of Birkhoffs ergodic theorem for Martin-Löf random points. Information and Computation 210, 21–30 (2012)

[8] Billingsley, P.: Probability and measure. John Wiley & Sons (2008)

[9] Birkhoff, G.D.: Proof of the ergodic theorem. Proceedings of the National Academy of Sciences 17(12), 656–660 (1931)

[10] Breiman, L.: The individual ergodic theorem of information theory. The Annals of Mathematical Statistics 28(3), 809–811 (1957)

[11] Cooper, S.B.: Computability theory. Chapman and Hall/CRC (2017)

[12] Downey, R.G., Hirschfeldt, D.R.: Algorithmic Randomness and Complexity. Theory and Applications of Computability, Springer New York, New York, NY (2010), http://dx.doi.org/10.1007/978-0-387-68441-3

[13] Ershov, Y.L.: A hierarchy of sets. I. Algebra and Logic 7(1), 25–43 (1968)

[14] Figueira, S., Hirschfeldt, D.R., Miller, J.S., Ng, K.M., Nies, A.: Counting the changes of random Δ_0^2 sets. Journal of Logic and Computation 25(4), 1073–1089 (2013)

[15] Franklin, J., Greenberg, N., Miller, J., Ng, K.M.: Martin-Löf random points satisfy Birkhoff's ergodic theorem for effectively closed sets. Proceedings of the American Mathematical Society 140(10), 3623–3628 (2012)

[16] Friedberg, R.M.: Two Recursively Enumerable Sets of Incomparable Degrees of Unsolvability (Solution of Post's Problem, 1944). Proceedings of the National Academy of Sciences of the United States of America 43(2), 236–238 (1957), http://www.jstor.org/stable/89817

[17] Granger, C.W.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society pp. 424–438 (1969)

[18] Hirschfeldt, D.R., Jockusch Jr, C.G., McNicholl, T.H., Schupp, P.E.: Asymptotic density and the coarse computability bound. Computability 5(1), 13–27 (2016)

[19] Hoeffding, W.: Probability inequalities for sums of bounded random variables. In: The Collected Works of Wassily Hoeffding, pp. 409–426. Springer (1994)

[20] Jockusch, C., Soare, R.: Degrees of members of Π_0^1 classes. Pacific Journal of Mathematics 40(3), 605–616 (1972)

[21] Jockusch Jr, C.G., Schupp, P.E.: Generic computability, Turing degrees, and asymptotic density. Journal of the London Mathematical Society 85(2), 472–490 (2012)

[22] Kalociński, D., Steifer, T.: An almost perfectly predictable process with no optimal predictor. In: 2019 IEEE International Symposium on Information Theory (ISIT). pp. 2504–2508. IEEE (2019)

[23] Kalociński, D., Steifer, T.: On unstable and unoptimal prediction. Mathematical Logic Quarterly 65(2), 218–227 (2019)

[24] Ko, K.I.: On the notion of infinite pseudorandom sequences. Theoretical Computer Science 48, 9–33 (1986)

[25] Kurtz, S.A.: Notions of weak genericity. The Journal of Symbolic Logic 48(3), 764–770 (1983)

[26] Martin-Löf, P.: The definition of random sequences. Information and Control 9(6), 602 – 619 (1966)

[27] Monin, B.: An answer to the Gamma question. In: Proceedings of the 33rd Annual ACM/IEEE Symposium on Logic in Computer Science. pp. 730–738. ACM (2018)

[28] Morvai, G., Yakowitz, S., Györfi, L., et al.: Nonparametric inference for ergodic, stationary time series. The Annals of Statistics 24(1), 370–379 (1996)

[29] Muchnik, A.A.: On the unsolvability of the problem of reducibility in the theory of algorithms. Dokl. Akad. Nauk SSSR 108, 194–197 (1956)

[30] Muchnik, A.A., Semenov, A.L., Uspensky, V.A.: Mathematical metaphysics of randomness. Theoretical Computer Science 207(2), 263–317 (1998)

[31] Nies, A.: Computability and randomness, vol. 51. OUP Oxford (2009)

[32] Ornstein, D.S.: Guessing the next output of a stationary process. Israel Journal of Mathematics 30(3), 292–296 (1978)

[33] Putnam, H.: Trial and error predicates and the solution to a problem of Mostowski. The Journal of Symbolic Logic 30(1), 49–57 (1965)

[34] Schnorr, C.P.: A unified approach to the definition of random sequences. Mathematical systems theory 5(3), 246–258 (Sep 1971), https://doi.org/10.1007/BF01694181

[35] Soare, R.I.: Turing computability. In: Theory and Applications of Computability. Springer (2016)

[36] Tadaki, K.: Phase transition and strong predictability. In: International Conference on Unconventional Computation and Natural Computation. pp. 340–352. Springer (2014)

[37] Takahashi, H.: On a definition of random sequences with respect to conditional probability. Information and Computation 206(12), 1375–1382 (2008)

[38] Tjur, T.: Conditional probability distributions. Institute of Mathematical Statistics, University of Copenhagen (1974)

[39] Turing, A.M.: On computable numbers, with an application to the Entscheidungsproblem. Proceedings of the London Mathematical Society 2(1), 230–265 (1937)

[40] Uspenskii, V.A., Semenov, A.L., Shen, A.K.: Can an individual sequence of zeros and ones be random? Russian Mathematical Surveys 45(1), 121 (1990)

[41] Van Lambalgen, M.: Von Mises' definition of random sequences reconsidered. The Journal of Symbolic Logic 52(3), 725–755 (1987)

[42] Ville, J.: Etude critique de la notion de collectif. Gauthier-Villars Paris (1939)