

Testing randomness online

Vladimir Vovk



практические выводы
теории вероятностей
могут быть обоснованы
в качестве следствий
гипотез о *предельной*
при данных ограничениях
сложности изучаемых явлений

On-line Compression Modelling Project (New Series)

Working Paper #24

First posted June 2, 2019. Last revised October 6, 2020.

Project web site:
<http://alrw.net>

Abstract

The hypothesis of randomness is fundamental in statistical machine learning and in many areas of nonparametric statistics; it says that the observations are assumed to be independent and coming from the same unknown probability distribution. This hypothesis is close, in certain respects, to the hypothesis of exchangeability, which postulates that the distribution of the observations is invariant with respect to their permutations. This paper reviews known methods of testing the two hypotheses concentrating on the online mode of testing, when the observations arrive sequentially. All known online methods for testing these hypotheses are based on conformal martingales, which are defined and studied in detail. An important variety of online testing is change detection, where the use of conformal martingales leads to conformal versions of the CUSUM and Shiryaev–Roberts procedures; these versions work in the nonparametric setting where the data is assumed IID according to a completely unknown distribution before the change. The paper emphasizes conceptual and practical aspects and states two kinds of results. Validity results limit the probability of a false alarm or, in the case of change detection, the frequency of false alarms for various procedures based on conformal martingales. Efficiency results establish connections between randomness, exchangeability, and conformal martingales.

Contents

1	Introduction	1
2	Conformal tests of randomness in the batch mode	3
3	Conformal martingales	6
4	Multistage nonrandomness detection	14
5	IID probability vs exchangeability probability	20
6	Conformal probability	23
7	Conclusion	26
	References	27
A	Connections with the algorithmic theory of randomness	32
B	Connections with the algorithmic theory of complexity	33

1 Introduction

A standard assumption in several areas of data science has been the assumption that the data is generated in the IID fashion, i.e., independently from the same distribution. This assumption is also known as the assumption of randomness (see, e.g., [64], [27, Section 7.1] and [58]). In this paper we are interested in testing this assumption.

The notion of randomness has been at the centre of discussions of the foundations of probability for at least 100 years, since Richard von Mises’s 1919 article [32]. For von Mises, random sequences (*collectives* in his terminology) served as the basis for probability theory and statistics, and other notions, such as probability, were defined in terms of collectives. Random sequences have been eclipsed in the foundations of mathematical probability theory by measure theory since Kolmogorov’s 1933 *Grundbegriffe* [19], but the conceptual side of randomness has been explored in the algorithmic theory of randomness (also initiated by Kolmogorov). We will discuss the conceptual side in Section 5 and then in Appendix A, but we start the main part of the paper with practical methods of detecting nonrandomness.

The most familiar mode of testing randomness in statistics is where we ask whether a given batch of data could have been generated by an IID process. In Section 2 we will see how the standard statistical tests for real-valued observations can be adapted to more general observation spaces.

We then move on to testing randomness online, assuming that observations arrive sequentially. Known methods of online testing of randomness are based on so-called conformal martingales. Conformal martingales are constructed on top of conventional machine-learning algorithms and have been used as a means of detecting deviations from randomness both in theoretical work (see, e.g., [58, Section 7.1], [15], [12]) and in practice (in the framework of the Microsoft Azure module on time series anomaly detection [66]). They provide an online measure of the amount of evidence found against the hypothesis of randomness: if the assumption of randomness is satisfied, a fixed nonnegative conformal martingale with a positive initial value is not expected to increase its initial value manyfold; on the other hand, if the hypothesis of randomness is seriously violated, a properly designed nonnegative conformal martingale with a positive initial value can be expected to increase its value substantially. Correspondingly, we have two desiderata for a nonnegative conformal martingale S :

- **Validity** is satisfied automatically: S is not expected to ever increase its initial value by much, under the hypothesis of randomness.
- But we also want to have **efficiency**, e.g., to have S_n/S_0 large with a high probability, if the hypothesis of randomness is violated.

Validity corresponds to controlling the error of the first kind in the language of statistical hypothesis testing, and efficiency is analogous to (but different from) statistical power (one minus the error of the second kind) in the Neyman–Pearson paradigm (see, e.g., [28]). In this paper a more Fisherian understanding of efficiency (introduced in Sections 5–6) will be more useful.

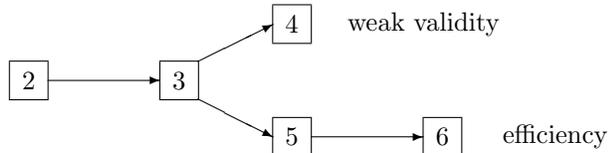


Figure 1: The structure of this paper (apart from the introduction and conclusion); the numbers refer to Sections 2–6.

Conformal martingales are defined and their validity is established in Section 3. Efficiency of a specific conformal martingale is not guaranteed and depends on the quality of the underlying machine-learning algorithm. Our exposition then branches in two directions, with Section 4 relaxing the requirement of validity and Sections 5–6 discussing efficiency; see Fig. 1.

It is often argued that the kind of validity enjoyed by nonnegative martingales is too strong, and we should instead be looking for a testing procedure that is valid only in the sense of not raising false alarms too often. In the context of testing randomness, the interpretation of the data-generating process adopted in Section 4 is that at first the data is IID, but starting from some moment T it ceases to be IID; the special case $T = 0$ describes the situation where the IID assumption is never satisfied (and so this interpretation does not restrict generality). We want our procedures to be efficient in the sense of raising an alarm soon after the null hypothesis (such as the assumption of randomness) becomes violated; both validity and efficiency can be required to hold with high probability or on average. In Section 4 conformal martingales are adapted to such less demanding requirements of validity using the standard CUSUM and Shiryaev–Roberts procedures.

Sections 5 and 6 discuss the much more difficult question of efficiency in the context of the strongly valid procedures of Section 3. We ask how much we can potentially lose when using conformal martingales as compared with unrestricted testing of either IID or exchangeability. We will see that at a crude scale customary in the algorithmic theory of randomness we do not lose much when restricting our attention to testing randomness with conformal martingales.

Our main running example will be the well-known USPS dataset of handwritten digits (see, e.g., [58, Section B.1]), which is known to be non-random. In Section 2 we check that it is really non-random: combining Bartels’s ratio test with the Nearest Neighbor method we obtain a tiny p-value. This fact, however, is not particularly useful. In Section 3 we report the performance of a nonnegative conformal martingale on the USPS dataset; it attains a huge final value. Such martingales are potentially very useful in practice allowing us to decide when a trained predictor needs to be retrained. The performance guarantees for prediction algorithms in mainstream machine learning are proven under the assumption of randomness, and after the deployment of such a predictor we

might want to monitor whether the new data remains IID. As soon as it ceases to be IID, it is wise to retrain the predictor.

We will also use a much less well-known dataset called “Absenteeism at work” (abbreviated to **Absenteeism**) and available at the UCI Machine Learning Repository [11]. The data was collected from July 2007 to July 2010 at a courier company in Brazil. We can imagine the management of the company monitoring the absences of the workforce. If the pattern of absences loses its stability (ceases to be IID), they might want to raise an alarm and investigate what is going on. In Section 3 we will construct a simple conformal martingale that finds decisive (to use Jeffreys’s [16, Appendix B] expression) evidence against the hypothesis of randomness for the dataset.

Connections with the algorithmic theory of randomness will be explained in Appendix A. The main part of this paper will not use the algorithmic notion of randomness; however, as customary in the algorithmic theory of randomness, in our discussions of efficiency we will concentrate on the binary case and on the case of a finite time horizon N . These restrictions go back to Kolmogorov (cf. [56, arXiv, Appendix A]); it would be interesting to eliminate them after a complete exploration of the binary and finite-horizon case (but we are not at that stage as yet).

Remark 1.1 (terminology related to the hypothesis of randomness). In this paper we discuss two main hypotheses about the data, randomness and exchangeability. The terminology related to the former is less standardized and will be summarized in this remark. We will use constantly (often adjectivally) two closely related terms, “randomness” and “IID”; the latter is, of course, always related to independent and identically distributed observations but is not always simply an abbreviation of “independent and identically distributed”. For example, *IID distributions* (or *IID measures*) are those of the form Q^N , where N is a natural number, or Q^∞ , Q being a probability measure. IID distributions generate IID observations. The hypothesis of randomness means that the data-generating distribution is an IID distribution (and so the observations are IID).

2 Conformal tests of randomness in the batch mode

As already mentioned, in this paper we are mainly interested in the *online* mode of testing: we assume that observations arrive sequentially, and after each observation we evaluate the degree to which the hypothesis of randomness has been discredited. We will discuss this online setting starting from the next section, but in this section discuss the *batch* setting, which is more standard in statistics.

Let us fix $N \in \mathbb{N} := \{1, 2, \dots\}$, the size of the batch. We would like to test the hypothesis of randomness using N observations. There are numerous standard tests of randomness in statistics: see, e.g., [27, Chapter 7]. These tests, however,

are usually applicable only to batches of real numbers, whereas in this paper we are interested in more general observations. In particular we will apply them to the USPS dataset of handwritten digits discussed in Section 1. To reduce the general case to real-valued observations we can apply basic ideas of conformal prediction [58]. Let \mathbf{Z} be a measurable space, the space of observations.

A *nonconformity measure* is a measurable function A mapping any finite sequence $(z_1, \dots, z_n) \in \mathbf{Z}^n$ of observations of any length $n \in \mathbb{N}$ to a sequence of numbers $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$ of the same length that is *equivariant* in the following sense: for any $n \in \mathbb{N}$ and any permutation $\pi : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$,

$$A(z_1, \dots, z_n) = (\alpha_1, \dots, \alpha_n) \implies A(z_{\pi(1)}, \dots, z_{\pi(n)}) = (\alpha_{\pi(1)}, \dots, \alpha_{\pi(n)}). \quad (1)$$

Intuitively, α_i (the *nonconformity score* of z_i) tells us how strange z_i looks as an element of the sequence (z_1, \dots, z_n) . (At this time the only relevant value is $n := N$, but in the next section we will need all $n \in \mathbb{N}$.)

Any conventional machine-learning algorithm can be turned (usually in more than one way) into a nonconformity measure. For example, suppose each observation z_i (an element of the USPS dataset) consists of two components, $z_i = (x_i, y_i)$, where $x_i \in [-1, 1]^{16 \times 16}$ is a hand-written digit (a 16×16 matrix of pixels, each pixel represented by its brightness on the scale $[-1, 1]$) and $y_i \in \{0, \dots, 9\}$ is its label (the true digit represented by the image). The 1-Nearest Neighbour algorithm can be turned into the nonconformity measure

$$\alpha_i := \frac{\min_{j \in \{1, \dots, n\}: y_j = y_i, j \neq i} d(x_i, x_j)}{\min_{j \in \{1, \dots, n\}: y_j \neq y_i} d(x_i, x_j)}, \quad (2)$$

where $d(x_i, x_j)$ is the Euclidean distance between x_i and x_j . Intuitively, a hand-written digit is regarded as strange if it is closer to a digit labeled in a different way than to digits labeled in the same way. See, e.g., [58, 1] for numerous other examples of nonconformity measures.

In our terminology we will follow [60]. A *p-variable for testing randomness* in Z^N , where Z is a measurable space, is a measurable function $f : Z^N \rightarrow [0, 1]$ such that, for any IID measure P on Z^N (i.e., for any measure P of the form Q^N , where Q is a probability measure on Z) and any $\epsilon > 0$,

$$P(\{(z_1, \dots, z_N) : f(z_1, \dots, z_N) \leq \epsilon\}) \leq \epsilon. \quad (3)$$

The value taken by f on the realized sample is then a bona fide p-value (perhaps conservative) for testing the hypothesis of randomness. Similarly, a *p-variable for testing exchangeability* in Z^N is a measurable function $f : Z^N \rightarrow [0, 1]$ such that, for any exchangeable probability measure P on Z^N (i.e., for any measure that is invariant with respect to permutations of observations) and any $\epsilon > 0$, we have (3). The values taken by such f are p-values for testing the hypothesis of exchangeability.

It is clear that every p-variable for testing exchangeability is a p-variable for testing randomness. Nonparametric statistics provides us with plenty of

p-variables for testing exchangeability in \mathbb{R}^N (under the rubric “testing of randomness”, even though they in fact test the weaker assumption of exchangeability; see, e.g., [64, 27, 2]). The following proposition shows how such a function f , in combination with a nonconformity measure A , generates a p-variable for testing exchangeability in \mathbf{Z}^N . Set

$$(f \circ A)(z_1, \dots, z_N) := f(A(z_1, \dots, z_N)).$$

Proposition 2.1. *If f is a p-variable for testing exchangeability in \mathbb{R}^N and A is a nonconformity measure, then $f \circ A$ is a p-variable for testing exchangeability in \mathbf{Z}^N .*

Proof. This follows immediately from the equivariance property (1): if P is an exchangeable probability measure on \mathbf{Z}^N , then its pushforward PA^{-1} is an exchangeable probability measure on \mathbb{R}^N , and so

$$P(f \circ A \leq \epsilon) = (PA^{-1})(f \leq \epsilon) \leq \epsilon.$$

To check that the pushforward PA^{-1} of an exchangeable probability measure P on \mathbf{Z}^N is indeed exchangeable, it suffices to notice that, for any permutation $\pi : \{1, \dots, N\} \rightarrow \{1, \dots, N\}$ and any event $E \subseteq \mathbf{Z}^N$,

$$\begin{aligned} & PA^{-1}(\{(u_1, \dots, u_N) : (u_{\pi(1)}, \dots, u_{\pi(N)}) \in E\}) \\ &= P(\{(z_1, \dots, z_N) : A(z_1, \dots, z_N) \in \{(u_1, \dots, u_N) : (u_{\pi(1)}, \dots, u_{\pi(N)}) \in E\}\}) \\ &= P(\{(z_1, \dots, z_N) : A(z_{\pi(1)}, \dots, z_{\pi(N)}) \in E\}) \\ &= P(\{(z_{\pi(1)}, \dots, z_{\pi(N)}) : A(z_{\pi(1)}, \dots, z_{\pi(N)}) \in E\}) = PA^{-1}(E). \end{aligned}$$

(The second equality follows from the equivariance of A and the third from the exchangeability of P .) \square

Example 2.2. Let us check that Proposition 2.1 ceases to be true if “exchangeability” is replaced by “randomness”. Suppose $\mathbf{Z} = [0, 1]$, and define the nonconformity score α_i of z_i in (z_1, \dots, z_n) by

$$\alpha_i := \begin{cases} 1 & \text{if } z_i \geq m_i \\ 0 & \text{otherwise,} \end{cases}$$

where m_i is the median of the multiset $\{z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n\}$. Suppose N is an even number (to simplify the notion of a median) and let $P := U^N$, where U is the uniform probability measure on $[0, 1]$. Then the pushforward PA^{-1} is concentrated on the subset of $\{0, 1\}^N$ containing equal numbers of 0s and 1s. (Roughly, this corresponds to half of the elements being above the median. Intuitively, A transforms a sequence that looks IID to a sequence that does not look IID at all for a large N , since it contains equal numbers of 0s and 1s.) By Stirling’s formula (see, e.g., (26) below), the probability of this subset is at most

$$\binom{N}{N/2} 2^{-N} \sim \sqrt{2/\pi} N^{-1/2} < N^{-1/2} \quad (4)$$

Table 1: Some p-values for the USPS dataset and Bartels’s ratio test

	Euclidean distance	Tangent distance
p-value	2.7×10^{-11}	7.5×10^{-16}

under any IID measure. Therefore, assuming N is large, the function f taking value $N^{-1/2}$ on this subset and 1 elsewhere on $\{0, 1\}^N$ is a p-variable for testing randomness while $f \circ A$ is not (indeed, $f \circ A$ will take value $N^{-1/2}$ almost surely under P , and then it is obvious that it cannot be a p-variable).

Table 1 gives the p-values produced by Bartels’s [2] ratio test applied to the nonconformity scores (2), where d is the Euclidean distance or the more sophisticated tangent distance [51, 18], as indicated. The p-values are very low, especially for the tangent distance.

Remark 2.3. It is important to keep in mind that the standard implementations of the tangent distance are not always symmetric and $d(x, x') \neq d(x', x)$ is possible (in particular, this is the case for Keyesers’s [18] implementation used in this paper). Whereas (2) itself defines a nonconformity measure regardless of the symmetry of d , efficient implementations of conformal testing of randomness based on (2) tend to rely on the symmetry of d and lose their validity if d is not symmetric. This is true for the implementation used for empirical studies in this paper; one possible solution (used here) is to replace $d(x_i, x_j)$ in (2) by the arithmetic mean of $d(x_i, x_j)$ and $d(x_j, x_i)$ (using the geometric mean produces similar results).

3 Conformal martingales

First let me give some basic definitions of conformal prediction (see, e.g., [58] or [1] for further details). Let us fix a nonconformity measure A . If Z is a set, Z^* is the set of all finite sequences of elements of Z ; if Z is a measurable space, Z^* is also regarded as a measurable space. The *p-value* p_n generated by A after being fed with a sequence of observations $(z_1, \dots, z_n) \in \mathbf{Z}^*$ is

$$p_n = p_n(z_1, \dots, z_n, \theta_n) := \frac{|\{i : \alpha_i > \alpha_n\}| + \theta_n |\{i : \alpha_i = \alpha_n\}|}{n}, \quad (5)$$

where i ranges over $\{1, \dots, n\}$, $\alpha_1, \dots, \alpha_n$ are the nonconformity scores for z_1, \dots, z_n ,

$$(\alpha_1, \dots, \alpha_n) = A(z_1, \dots, z_n),$$

and θ_n is a random number distributed uniformly on the interval $[0, 1]$. The following proposition gives the standard property of validity for conformal prediction (for a proof, see, e.g., [58, Proposition 2.8]).

Proposition 3.1. *Suppose the observations z_1, z_2, \dots are IID, $\theta_1, \theta_2, \dots$ are IID and distributed uniformly on $[0, 1]$, and the sequences z_1, z_2, \dots and $\theta_1, \theta_2, \dots$ are independent. Then the p-values p_1, p_2, \dots as defined in (5) are IID and distributed uniformly on $[0, 1]$.*

Remark 3.2. On a few occasions, we will also need a version of Proposition 3.1 for a finite horizon $N \in \mathbb{N}$. Now we have finite input sequences z_1, \dots, z_N and $\theta_1, \dots, \theta_N$ and, correspondingly, a finite output sequence p_1, \dots, p_N . The conclusion of Proposition 3.1 will still hold even if we relax the assumption of z_1, \dots, z_N being IID to the assumption that they are exchangeable. (See [58, Theorem 8.2].)

The formal definition of a nonnegative conformal martingale (equivalent to the definition given in [58, Section 7.1]) given in this paragraph will be followed by a discussion of the intuition behind it in the following paragraph (in our informal discussions we will often abbreviate “nonnegative conformal martingale” to “conformal martingale”). A *betting martingale* is a measurable function $F : [0, 1]^* \rightarrow [0, \infty]$ such that, for each sequence $(u_1, \dots, u_{n-1}) \in [0, 1]^{n-1}$, $n \geq 1$, we have

$$\int_0^1 F(u_1, \dots, u_{n-1}, u) du = F(u_1, \dots, u_{n-1}); \quad (6)$$

notice that betting martingales are required to be nonnegative. A *nonnegative conformal martingale* is any sequence of functions $S_n : (\mathbf{Z} \times [0, 1])^\infty \rightarrow [0, \infty]$, $n = 0, 1, \dots$, such that, for some nonconformity measure A and betting martingale F , for all $m \in \{0, 1, \dots\}$, $(z_1, z_2, \dots) \in \mathbf{Z}^\infty$, and $(\theta_1, \theta_2, \dots) \in [0, 1]^\infty$,

$$S_m(z_1, \theta_1, z_2, \theta_2, \dots) = F(p_1, \dots, p_m), \quad (7)$$

where p_n , $n \in \mathbb{N}$, is the p-value computed by (5) from the nonconformity measure A , the observations z_1, z_2, \dots , and the n th element θ_n of the sequence $(\theta_1, \theta_2, \dots)$. Notice that $S_m(z_1, \theta_1, z_2, \theta_2, \dots)$ depends on $z_1, \theta_1, z_2, \theta_2, \dots$ only via $z_1, \theta_1, \dots, z_m, \theta_m$.

Intuitively, a betting martingale describes the evolution of the capital of a player who gambles against the hypothesis that the p-values p_1, p_2, \dots are distributed uniformly and independently, as they should under the hypothesis of randomness (see Proposition 3.1). The requirement (6) expresses the fairness of the game: at step $n - 1$, the conditional expected value of the player’s future capital at step n given the present situation (i.e., the first $n - 1$ p-values) is equal to his current capital. This formalization of fair betting goes back to Ville [54] and was made very popular in probability theory by Doob [9]; for a recent review of developments in various directions, see [44]. A conformal martingale is what we get when we feed a betting martingale with p-values (5) produced by conformal prediction.

One way of constructing betting martingales is to use “betting functions”, in the terminology of [12]. A *betting function* $f : [0, 1] \rightarrow [0, \infty]$ is a function satisfying $\int_0^1 f(u) du = 1$. A useful method of betting against the hypothesis that

the p-values p_1, p_2, \dots are independent and uniformly distributed is to choose, before each step n , a betting function f_n that may depend on p_1, \dots, p_{n-1} (in a measurable manner). Then

$$F(p_1, \dots, p_n) := f_1(p_1) \dots f_n(p_n), \quad n = 0, 1, \dots, \quad (8)$$

will be a betting martingale (a conformal martingale if p_1, p_2, \dots are generated by conformal prediction, (5)).

Remark 3.3. Conformal martingales are *exchangeability martingales*, i.e., stochastic processes that are martingales with respect to any exchangeable distribution. The existence of non-trivial exchangeability martingales is, however, not obvious. It is easy to check that for the natural underlying filtration $(\mathcal{F}_n)_{n=0,1,\dots}$ generated by the observations z_1, z_2, \dots the only exchangeability martingales are almost sure constants. There are two reasons why non-trivial exchangeability martingales exist:

- Our underlying filtration is poorer than \mathcal{F}_n . A conformal martingale S satisfies

$$\mathbb{E}(S_n \mid S_1, \dots, S_{n-1}) = S_{n-1}, \quad n \in \mathbb{N},$$

i.e., it is a martingale in its own filtration. Moreover, it is a martingale in the filtration $(\mathcal{G}_n)_{n=0,1,\dots}$ where \mathcal{G}_n is generated by the first n p-values p_1, \dots, p_n .

- Conformal martingales are randomized: they also depend on the random numbers $\theta_1, \theta_2, \dots$.

The first reason alone seems to be insufficient for getting really useful exchangeability martingales: e.g., in the binary case $\mathbf{Z} = \{0, 1\}$, the observations z_1, \dots, z_n are determined by the values S_0, S_1, \dots, S_n , unless $S_i = S_{i-1}$ for some $i \in \{1, \dots, n\}$ (let us check this for $n = 1$: depending on the value of z_1 , we have either $S_1(z_1, \dots) > S_0$ or $S_1(z_1, \dots) < S_0$, and knowing which inequality is true determines z_1 ; for general n , use induction in n). In many practically interesting cases there is not much randomness in conformal martingales; it is only used for tie-breaking. However, even a tiny amount of randomness can be conceptually important (other fields where this phenomenon has been observed are differential privacy and defensive forecasting [44, Section 12.7]).

Remark 3.4. Notice that exchangeability martingales discussed in Remark 3.3 can be equivalently defined as stochastic processes that are martingales with respect to any IID distribution, assuming that the observation space \mathbf{Z} is a Borel space (this is a very weak requirement; see, e.g., [39, B.3.2]). Indeed, according to de Finetti's theorem (see, e.g., [39, Theorem 1.49]) every exchangeable distribution is then a Bayesian mixture of IID distributions, and so being a martingale with respect to all IID distributions and with respect to all exchangeable distributions are equivalent.

Using conformal martingales for testing randomness

We only consider nonnegative conformal martingales S with $S_0 \in (0, \infty)$. Let us see how such martingales can be used for testing randomness.

A possible goal is to raise an alarm warning the user about lack of randomness as soon as possible. Ville’s inequality [48, Chapter 7, Section 3, Theorem 1.III] says that, for any $c > 1$,

$$\mathbb{P}(\exists n : S_n/S_0 \geq c) \leq 1/c$$

under any IID distribution. This means that if we raise an alarm when S_n/S_0 reaches threshold c , we will be wrong with probability at most $1/c$. This is a strong (in some situations too strong) requirement of validity, and we will sometimes refer to it as *strong validity*.

We can also interpret S_n/S_0 directly as the amount of evidence detected against the first n observations being IID. In principle, there is no need to raise an alarm explicitly, and we can leave the decision whether to abandon the assumption of randomness with the user of our methods.

As an example, for the USPS dataset of handwritten digits (9298 in total), the online performance of a nonnegative conformal martingale based on the nonconformity measure (2) (with Euclidean distance) is shown in the left panel of Fig. 2 (which is Fig. 7.6 in [58], where full details of the conformal martingale can be found). We already know from Section 2 that the USPS dataset is not random, and its lack of randomness is detected by this conformal martingale in the online mode. The advantage of the online mode is that such a conformal martingale may be used in practice for deciding when a digit classifier needs to be retrained (we can see that approximately after the 2400th observation it would be definitely desirable).

The conformal martingale whose performance is shown in Fig. 2 is fairly complicated, but the ideas behind its construction are instructive. Any function

$$f^{(\kappa)}(u) := \kappa u^{1-\kappa}, \quad u \in [0, 1], \quad \kappa \in (0, 1), \quad (9)$$

is a betting function, in the sense of satisfying $\int_0^1 f^{(\kappa)}(u) du = 1$. This implies that

$$F^{(\kappa)}(u_1, \dots, u_n) := \prod_{i=1}^n f^{(\kappa)}(u_i)$$

is a betting martingale (satisfies (6)) and so determines, by (7), a conformal martingale $S = S^{(\kappa)}$. To get rid of the dependence on κ , we can use the conformal martingale $\int_0^1 S^{(\kappa)} d\kappa$, which was called the *simple mixture* in [58, Section 7.1]. The simple mixture starts from 1 and its final value is 2.18×10^{10} .

The simple mixture attains an astronomical final value, but it can be improved further, in some sense tracking the best value of κ (following Herbster and Warmuth’s [14] idea of tracking the best expert). For a stochastic process producing a random sequence $\kappa_1, \kappa_2, \dots$, we can integrate

$$F^{(\kappa_1, \kappa_2, \dots)}(u_1, \dots, u_n) := \prod_{i=1}^n f^{(\kappa_i)}(u_i)$$

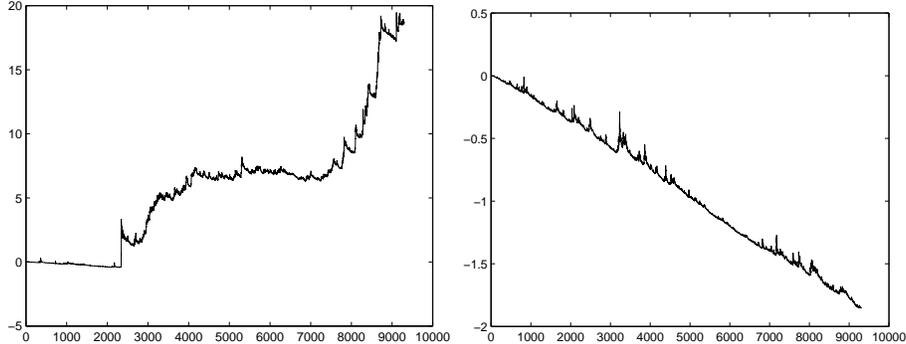


Figure 2: Left panel: The values S_n of a nonnegative conformal martingale after observing the first n digits, $n = 0, \dots, 9298$, of the USPS dataset, with the log-10 scale for the vertical axis. The initial value S_0 is 1, and the final value S_{9298} is 4.71×10^{18} . Right panel: The values S_n of the same nonnegative conformal martingale as in the left panel after observing the first n digits of a randomly permuted USPS dataset, with the log-10 scale for the vertical axis. The initial value S_0 is 1, and the final value S_{9298} is 0.0142.

with respect to the distribution of $\kappa_1, \kappa_2, \dots$, and for a reasonable choice of the stochastic process, we can improve the final value of the conformal martingale. It can be further boosted by allowing the stochastic process to “sleep” at some steps (so that the corresponding conformal martingale does not gamble on those steps). In statistical literature, these ideas are discussed in [53].

The idea behind the betting functions (9) is that the lack of randomness will show itself in abnormally low p-values. Fedorova and Nourtdinov [12] came up with an unexpected new idea: in fact, we can gamble against *any* non-uniformity in the distribution of the p-values, and this may be a very successful strategy for detecting non-randomness.

Suppose we know the true distribution of the n th p-value p_n (conditional on knowing the first $n - 1$ p-values), and suppose it is continuous with density ρ . What betting function f should we choose? This is a continuous version of the standard problem of horse race betting [17, 8]. The following simple lemma sheds some light on ways of exploiting such non-uniformity.

Lemma 3.5. *For any probability density functions ρ and f on $[0, 1]$ (so that $\int_0^1 \rho(p) dp = 1$ and $\int_0^1 f(p) dp = 1$),*

$$\int_0^1 (\log \rho(p)) \rho(p) dp \geq \int_0^1 (\log f(p)) \rho(p) dp \quad (10)$$

and

$$\int_0^1 (\log \rho(p)) \rho(p) dp \geq 0. \quad (11)$$

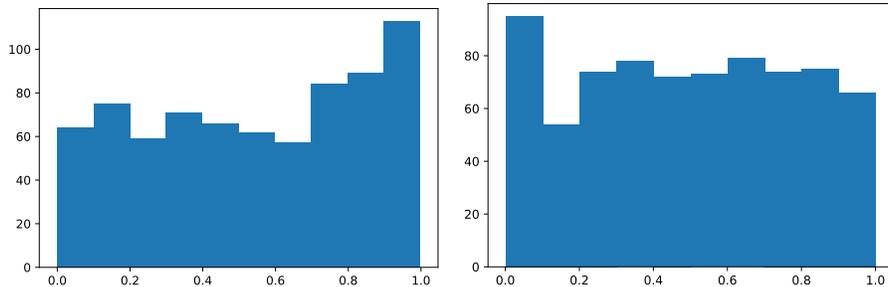


Figure 3: Left panel: The histogram (with 10 bins) of the p-values at the last step for the nonnegative conformal martingale of Fig. 4 on the **Absenteeism** dataset. Right panel: The p-values of the same nonnegative conformal martingale on the randomly permuted **Absenteeism** dataset.

Proof. It is well known (and immediately follows from the inequality $\log x \leq x - 1$) that the Kullback–Leibler divergence is always non-negative:

$$\int_0^1 \log\left(\frac{\rho(p)}{f(p)}\right) \rho(p) dp \geq 0.$$

This is equivalent to (10). And (11) is a special case of (10) corresponding to the probability density function $f := 1$. \square

If we choose a betting function f , the log of our capital will increase by the right-hand side of (10) in expectation. Therefore, according to (10), the largest increase in expectation is achieved when we use ρ as the betting function. (Increasing the log capital as much as possible in expectation is a natural objective since such increases add to give the log of the final capital, and so we can apply the law of large numbers, as in horse racing [8, Section 6.1] or log-optimal portfolios [8, Chapter 16].) The discrete version of this strategy is known as *Kelly gambling* [8, Theorem 6.1.2].

How efficient the betting function ρ depends on the left-hand side of (10), which is the minus (differential) entropy of ρ . The maximum entropy distribution on $[0, 1]$ is the uniform distribution, as asserted by (11), whose right-hand side is equal to the minus entropy of the uniform distribution. (This is a very special case of standard maximum entropy results, such as [8, Theorem 12.1.1].) The uniform true distribution for the p-values gives zero expected increase in the log capital; otherwise, it is positive.

Let us apply these ideas to the **Absenteeism** dataset briefly described in Section 1. We will estimate the distribution of the past p-values using a histogram and then will use the estimated distribution for betting (therefore, we will implicitly assume the stability of the distribution of p-values). The dataset consists of 740 observations (employees’ absences) and a number of attributes; it is given in the Clustering section of the repository, and so there is no specified

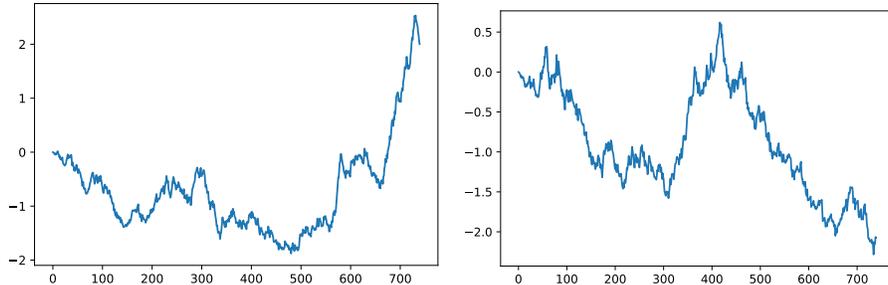


Figure 4: Left panel: The values S_n of a nonnegative conformal martingale after observing the first n absences, $n = 0, \dots, 740$, in the **Absenteeism** dataset, with the log-10 scale for the vertical axis. The initial value S_0 is 1, and the final value S_{740} is 100.50. Right panel: The values S_n of the same nonnegative conformal martingale as in the left panel after observing the first n absences in a randomly permuted **Absenteeism** dataset, with the log-10 scale for the vertical axis. The initial value S_0 is 1, and the final value S_{740} is 0.00841.

label. Let us use four attributes, Age, Disciplinary failure, Education, and Son (meaning the number of children); the other attributes appeared to me more subjective (such as Social drinker and Social smoker) or less relevant. To apply the same nonconformity measure as before, (2), we need to nominate one of the attributes as label, and “Disciplinary failure” appears to be particularly relevant for studying the phenomenon of absenteeism. To make the attributes comparable, let us divide Age by 50, Education by 3 (this attribute ranges from 1, high school, to 4, master and doctor), and Son by 4. With this choice, the histogram of p-values at the last step is given in the left panel of Fig. 3, and it is visibly non-uniform, namely tends to increase from left to right (unlike the more stable right panel showing the p-values for a randomly permuted dataset; cf. Remark 3.2).

Figure 4 shows the results for a simple conformal martingale exploiting the non-uniformity of the p-values. We maintain B bins corresponding to the subintervals $I_i := [(i-1)/B, i/B]$ of $[0, 1]$, $i = 1, \dots, B$ (we ignore the possibility of p-values landing on a boundary between two bins). Initially, each bin contains C p-values (these dummy p-values are an element of regularization). At step n , after computing the n p-values for this step (using the nonconformity measure (2)), the algorithm puts them into the corresponding bins and uses as its betting function f_n the function equal to $(C + n_i)/(C + n/B)$ on the i th bin, where n_i is the number of p-values in that bin. The mean of the betting function is 1 (and $C + n/B$ is the normalizing constant), and it approximates Kelly gambling. The conformal martingale whose performance is shown in Fig. 4 is (8); its initial value is 1.

The expectation of the martingale’s final value is 1 (it is an *e-value* [61, 13, 41]), and we can use Jeffreys’s [16, Appendix B] rule of thumb for interpret-

ing the amount of evidence against the null hypothesis of randomness that it provides:

- A value below 1 supports the null hypothesis.
- A value in the interval $(1, \sqrt{10})$ provides poor evidence against the null hypothesis (is not worth more than a bare mention).
- A value in $(\sqrt{10}, 10)$ provides substantial evidence.
- For a value in $(10, 10^{3/2})$, the evidence is strong.
- For a value in $(10^{3/2}, 100)$, the evidence is very strong.
- Finally, for values above 100 the evidence is decisive.

For example, the evidence against the hypothesis of randomness provided by the left-hand panel of Fig. 4, which uses $B = C = 10$, is decisive (although bordering on very strong) since the final value of the conformal martingale is 100.50. Varying various parameters leads to similar, often stronger, results. For example, replacing the ratio in (2) by difference, adding two extra attributes Social drinker and Social smoker, and setting $B = C = 20$, leads to a final value of 3446.75 greatly exceeding the threshold for decisive evidence. However, all of these conformal martingales find deviations from randomness only at the end of the period (starting from the 650th observation at the earliest).

Optimality of sequential testing procedures

The efficiency of our procedure for testing randomness will be the topic of Section 6, and in this section we will only discuss the nature of the problem. Perhaps the most satisfactory results about the efficiency of sequential testing procedures are optimality results such as that obtained by Wald and Wolfowitz [65] for Wald’s [62, 63] sequential probability ratio test. The goal of establishing such optimality results for our procedures for testing randomness would be, however, too ambitious.

Wald’s sequential probability ratio test was designed by him in April 1943 [62, Section B] for the problem of testing a simple hypothesis against a simple alternative, with IID observations. Let S_n be the likelihood ratio of the alternative hypothesis to the null hypothesis after n observations. The test consists in fixing two positive constants A and B such that $A > 1 > B$ and stopping as soon as S_n leaves the interval (B, A) . If $S_n > A$ at that time n , we reject the null hypothesis; otherwise, we accept it.

A sequential test can make errors of two kinds: reject the null hypothesis when it is true (error of the first kind) or accept it when it is false (error of the second kind). Any sequential probability ratio test T is efficient in a strong sense: if another sequential test T' has errors of the first and second kind that are not worse than those for T , the expected time of reaching a decision is as good for T as it is for T' (or better), under both null and alternative hypotheses. In other

words, sequential probability ratio tests optimize the number of observations needed to arrive at a decision, under natural constraints.

Wald showed the efficiency of his test in the sections “Efficiency of the Sequential Probability Ratio Test” in [62, 63] ignoring the possibility of S_n overshooting A or undershooting B . In [65] he and Wolfowitz provided a full proof.

The strength of this result is made possible by the restricted nature of the testing problem. Both null and alternative hypotheses are known probability distributions. The test is specified by two numbers, A and B . The situation with testing randomness using conformal martingales is very different. A conformal martingale is determined by the underlying nonconformity measure, which can even involve an element of intelligence. See, e.g., [58], which defines numerous nonconformity measures based on powerful algorithms of machine learning, including neural networks. We cannot expect to be able to *prove* that such a procedure is successful (as argued by philosophers in other contexts; see, e.g., [36, Section 20]). The task of designing a conformal martingale is too open-ended for that.

Our approach to establishing the efficiency of our strongly valid procedures will not be based on optimality. The idea is to show that our procedures do not constrain us: whatever a procedure for testing randomness can achieve, can be achieved with conformal martingales. Notice that even the Wald–Wolfowitz result can be interpreted in this way. However, our results in Section 6 will be much cruder.

4 Multistage nonrandomness detection

Our main concern in this section is application of conformal prediction to online change detection, which we already started discussing in Section 1. A typical example of online change detection is where we observe attacks, which we assume to be IID, on a computer system. When a new kind of attacks appears, the process of attacks ceases to be IID, and we would like to raise an alarm soon afterwards. The two benchmark datasets that we considered in the previous section can also be used to illustrate the problem of change detection: we may be interested in deciding when to retrain a predictor and in detecting a change in the pattern of workforce absences.

There is vast literature on online change detection; see, e.g., [35, 49] for reviews. However, the standard case is where the pre-change and post-change distributions are known, and the only unknown is the time of change. Generalizations of this picture usually stay fairly close to it (see, e.g., [35, Section 7.3]). Conformal change detection relaxes the standard assumptions radically.

Remark 4.1. The literature on batch change detection is also vast; see, e.g., [5] for an early review. Here the problem is to detect changes in a data sequence all of which is given to us in a batch rather than sequentially. The importance of this problem has grown in recent decades because of its applications in bioinformatics; see, e.g., [50]. This paper, however, concentrates on online problems.

As explained in Section 1, we may regard any problem of detecting non-randomness in the online mode as a problem of detecting a change point. The latter includes as special case the situation where the assumption of randomness is never satisfied, since 0 is an allowed change point. Our informal goal is to raise an alarm as soon as possible after the hypothesis of randomness ceases to be true. In the previous section we did not insist on having an explicit rule for raising an alarm, and simply regarded the value of a nonnegative conformal martingale starting from 1 as the amount of evidence found against the hypothesis of randomness, but in this section it will be more convenient to couch our discussion in terms of such rules.

As already mentioned, the kind of guarantees provided by the policy of raising an alarm when $S_n/S_0 \geq c$ is often regarded as too strong to be really useful. This can be illustrated using the analogue of the left panel of Fig. 2 for a randomly permuted USPS dataset. The same conformal martingale performs as shown in the right panel of Fig. 2 (this is Fig. 7.8 in [58]). The conformal martingale is trying to gamble against an exchangeable sequence of observations, which is futile, and so its value decreases exponentially quickly. If a change occurs at some point in the distant future, it might take a long time for the martingale to recover its value. This is a general phenomenon; we must pay for giving ourselves the chance to detect lack of exchangeability by losing capital in the situation of exchangeability.

The conformal martingale in Fig. 2 is based on the ideas of tracking the best expert and sleeping, which often make it easier to recover the martingale’s value (to “catch up”), as demonstrated by van Erven et al. [53]. The right panel of Fig. 2 suggests that this approach has its limits, and we need to do something more radical: change the rules of the game.

Weaker guarantees are provided by multistage procedures originated, in a basic form, by Shewhart in his control chart techniques [45] and perfected by Page [33] and Kolmogorov and Shiryaev [25]. As Shiryaev mentions in his fascinating historical account [47, Section 1], he and Kolmogorov rejected the policy of raising an alarm when $S_n/S_0 \geq c$ in favor of a multistage procedure, which was “the *correct formulation of the problem*” (the emphasis is Shiryaev’s), in January 1959 or soon afterwards, after talking to a practitioner, Yurii B. Kobzarev, the founder of the Soviet school of radiolocation.

CUSUM-type change detection

A standard multistage procedure of raising alarms is the CUSUM procedure proposed by Page [33] (see also [35, Section 6.2]). According to this procedure, we raise the k th alarm at the time

$$\tau_k := \min \left\{ n > \tau_{k-1} : \max_{i=\tau_{k-1}, \dots, n-1} \frac{S_n}{S_i} \geq c \right\}, \quad k \in \mathbb{N}, \quad (12)$$

where the threshold $c > 1$ is a parameter of the algorithm, $\tau_0 := 0$, and $\min \emptyset := \infty$. If $\tau_k = \infty$ for some k , an alarm is raised only finitely often; otherwise it is raised infinitely often. The procedure is usually applied to the

likelihood ratio process between two IID distributions, but it can be applied to any positive martingale, and in this paper we are interested in the case where S is a conformal martingale, which is now additionally assumed to be positive, ensuring that the denominator in (12) is always non-zero. CUSUM is often interpreted as a repeated sequential probability ratio test [33, Section 4.2]. The *conformal CUSUM procedure* (i.e., CUSUM applied to a positive conformal martingale) was introduced in [55]; however, a basic and approximate version of this procedure has been known since 1990: see [31].

Properties of validity for the conformal CUSUM procedure will be obtained in this paper as corollaries of the corresponding properties of validity for the Shiryaev–Roberts procedure, which we consider next.

Shiryaev–Roberts change detection

A popular alternative to the CUSUM procedure is the Shiryaev–Roberts procedure [46, 38], which modifies (12) as follows:

$$\tau_k := \min \left\{ n > \tau_{k-1} : \sum_{i=\tau_{k-1}}^{n-1} \frac{S_n}{S_i} \geq c \right\}, \quad k \in \mathbb{N} \quad (13)$$

(i.e., we just replace the max in (12) by \sum). We will again apply it to a conformal martingale S , still assumed to be positive, obtaining the *conformal Shiryaev–Roberts procedure*.

The procedure defining τ_1 is based on the statistics

$$R_n := \sum_{i=0}^{n-1} \frac{S_n}{S_i}, \quad (14)$$

which admit the recursive representation

$$R_n = \frac{S_n}{S_{n-1}} (R_{n-1} + 1), \quad n \in \mathbb{N}, \quad (15)$$

with $R_0 := 0$. An interesting finance-theoretic interpretation of this representation is that R_n is the value at time n of a portfolio that starts from \$0 at time 0 and invests \$1 into the martingale S at each time $i = 1, 2, \dots$ [10, Section 2]. If and when an alarm is raised at time n , we apply the same procedure to the remaining observations z_{n+1}, z_{n+2}, \dots .

The following proposition gives a non-asymptotic property of validity of the Shiryaev–Roberts procedure. Roughly, it says that we do not expect the first alarm to be raised too soon under the hypothesis of randomness.

Proposition 4.2. *The conformal Shiryaev–Roberts procedure (13) satisfies $\mathbb{E}(\tau_1) \geq c$, for any $c > 1$, under the assumptions of Proposition 3.1.*

Of course, we can apply Proposition 4.2 to other alarm times as well obtaining $\mathbb{E}(\tau_k - \tau_{k-1}) \geq c$ for all $k \in \mathbb{N}$ (and similar inequalities for some conditional

expectations, as discussed below in the proof of Proposition 4.4). Therefore, more generally, the time interval between raising successive alarms is not too short under the hypothesis of randomness.

All results of this section (from Proposition 4.2 to Corollary 4.6) are general and applicable to any positive martingale S . However, they are usually stated for S being the likelihood ratio between two IID distributions (pre-change and post-change). To simplify exposition, I will state them only for S being a positive conformal martingale with the underlying filtration $(\mathcal{G}_n)_{n=0,1,\dots}$, where \mathcal{G}_n is generated by the first n p-values p_1, \dots, p_n . However, our arguments (which are standard in literature on change detection) will be applicable to any filtration and any positive martingale with respect to that filtration.

Proof of Proposition 4.2. The proof will follow from the fact that $R_n - n$ is a martingale; this fact (noticed, in a slightly different context, in [34, Theorem 1]) follows from (15): since S is a martingale,

$$\mathbb{E}(R_n | \mathcal{G}_{n-1}) = \frac{\mathbb{E}(S_n | \mathcal{G}_{n-1})}{S_{n-1}} (R_{n-1} + 1) = R_{n-1} + 1.$$

Another condition for $R_n - n$ being a martingale requires the integrability of R_n , which follows from the integrability of each addend in (14):

$$\mathbb{E}\left(\frac{S_n}{S_i}\right) = \mathbb{E}\left(\mathbb{E}\left(\frac{S_n}{S_i} | \mathcal{G}_i\right)\right) = \mathbb{E}(1) = 1 < \infty.$$

Fix the threshold $c > 1$. By Doob's optional sampling theorem (see, e.g., [48, Chapter 7, Section 2, Theorem 1]) applied to the martingale $R_n - n$,

$$\mathbb{E}(\tau_1) = \mathbb{E}(R_{\tau_1}) \geq c.$$

Applying this theorem, however, requires some regularity conditions, and the rest of this proof is devoted to checking technical details.

If $\tau_1 = \infty$ with a positive probability, we have $\mathbb{E}(\tau_1) = \infty \geq c$, and so we assume that $\tau_1 < \infty$ a.s. Doob's optional sampling theorem is definitely applicable to the stopping time $\tau_1 \wedge L$, where L is a positive constant (see, e.g., [48, Chapter 7, Section 2, Corollary 1]), and so the nonnegativity of R implies

$$\mathbb{E}(\tau_1) \geq \mathbb{E}(\tau_1 \wedge L) = \mathbb{E}(R_{\tau_1 \wedge L}) \geq \mathbb{E}(R_{\tau_1} 1_{\{\tau_1 \leq L\}}) \geq c \mathbb{P}(\tau_1 \leq L) \rightarrow c$$

as $L \rightarrow \infty$, 1_E being the indicator function of an event E . \square

Corollary 4.3. *The conformal CUSUM procedure (12) also satisfies $\mathbb{E}(\tau_1) \geq c$ under the assumptions of Proposition 3.1.*

Proof. All our properties of validity for the CUSUM procedure will be deduced from the corresponding properties for Shiryaev–Roberts and the fact that Shiryaev–Roberts raises alarms more often than CUSUM does, in the following sense. Let τ_k (resp. τ'_k) be the time of the k th alarm raised by Shiryaev–Roberts (resp. CUSUM). Then $\tau_k \leq \tau'_k$ for all k ; this can be checked by induction in k . \square

The next proposition is an asymptotic counterpart of Proposition 4.2 given in terms of frequencies.

Proposition 4.4. *Let A_n be the number of alarms*

$$A_n := \max\{k : \tau_k \leq n\}$$

raised by the conformal Shiryaev–Roberts procedure (13) after seeing the first n observations z_1, \dots, z_n . Then, under the assumptions of Proposition 3.1,

$$\limsup_{n \rightarrow \infty} \frac{A_n}{n} \leq \frac{1}{c} \quad a.s. \quad (16)$$

Under the assumptions of Proposition 3.1, all alarms are false, and so (16) limits the frequency of false alarms.

Proof. Fix a positive conformal martingale S and a threshold $c > 0$. We can rewrite (13) as

$$\tau_k := \min\{n > \tau_{k-1} : R_n^k \geq c\}, \quad (17)$$

where

$$R_n^k := \sum_{i=\tau_{k-1}}^{n-1} \frac{S_n}{S_i}.$$

It will be convenient to modify (17) by forcing an alarm L steps after the last one:

$$\tau'_k := (\tau'_{k-1} + L) \wedge \min\{n > \tau'_{k-1} : R_n'^k \geq c\},$$

where $\tau'_0 := 0$ and

$$R_n'^k := \sum_{i=\tau'_{k-1}}^{n-1} \frac{S_n}{S_i}.$$

(The value of L will be chosen later.) Similarly to the proof of Corollary 4.3, by induction in k we can check that, for all k , $\tau'_k \leq \tau_k$.

We still have a recursive representation similar to (15) for (R^k) and (R'^k) . Notice that $R_n'^k$, $n \geq \tau'_{k-1}$, is a nonnegative submartingale with $n - \tau'_{k-1}$ as its compensator (and we can set $R_n'^k$ and its compensator to 0 for $n < \tau'_{k-1}$).

Remember that \mathcal{G}_n is the σ -algebra generated by the p-values p_1, \dots, p_n , and let $\mathcal{G}_{\tau'_k}$ be the σ -algebra of events E such that $E \cap \{\tau'_k \leq n\} \in \mathcal{G}_n$ for all n (informally, $\mathcal{G}_{\tau'_k}$ consists of the events E expressible in terms of the p-values and settled at time τ'_k).

Let us say that $k \in \mathbb{N}$ is *slow* if

$$\mathbb{P}\left(\tau'_k - \tau'_{k-1} = L \mid \mathcal{G}_{\tau'_{k-1}}\right) \geq c/L;$$

otherwise, k is *fast*. Notice that the event that k is fast (or slow) is $\mathcal{G}_{\tau'_{k-1}}$ -measurable. By Doob's optional sampling theorem and the nonnegativity of $R_n'^k$, where $n \geq \tau'_{k-1}$, for a fast k we obtain, similarly to the proof of Proposition 4.2,

$$\mathbb{E}\left(\tau'_k - \tau'_{k-1} \mid \mathcal{G}_{\tau'_{k-1}}\right) = \mathbb{E}\left(R_{\tau'_k}^k \mid \mathcal{G}_{\tau'_{k-1}}\right)$$

$$\begin{aligned}
&= \mathbb{E} \left(R'_{\tau'_k} 1_{\{\tau'_k - \tau'_{k-1} = L\}} \mid \mathcal{G}_{\tau'_{k-1}} \right) + \mathbb{E} \left(R'_{\tau'_k} 1_{\{\tau'_k - \tau'_{k-1} < L\}} \mid \mathcal{G}_{\tau'_{k-1}} \right) \\
&\geq 0 + c \mathbb{E} \left(1_{\{\tau'_k - \tau'_{k-1} < L\}} \mid \mathcal{G}_{\tau'_{k-1}} \right) \geq c(1 - c/L) = c - c^2/L.
\end{aligned}$$

Let $F \subseteq \mathbb{N}$ be the random set of all fast k , $S := \mathbb{N} \setminus F$ be the random set of all slow k , and F_K (resp. S_K) be the set consisting of the K smallest elements of F (resp. S). The strong law of large numbers for bounded martingale differences now implies

$$\liminf_{K \rightarrow \infty} \frac{1}{K} \sum_{k \in S_K} (\tau'_k - \tau'_{k-1}) \geq L(c/L) = c \quad \text{a.s.} \quad (18)$$

and

$$\liminf_{K \rightarrow \infty} \frac{1}{K} \sum_{k \in F_K} (\tau'_k - \tau'_{k-1}) \geq c - c^2/L \quad \text{a.s.}; \quad (19)$$

the inequality in (18) (resp. (19)) is interpreted as true when $|S| < \infty$ (resp. $|F| < \infty$). Combining (18) and (19), we obtain

$$\liminf_{K \rightarrow \infty} \frac{\tau'_K}{K} = \liminf_{K \rightarrow \infty} \frac{1}{K} \sum_{k=1}^K (\tau'_k - \tau'_{k-1}) \geq c - c^2/L \quad \text{a.s.}$$

Therefore, setting

$$A'_n := \max\{k : \tau'_k \leq n\},$$

we have

$$\limsup_{n \rightarrow \infty} \frac{A_n}{n} \leq \limsup_{n \rightarrow \infty} \frac{A'_n}{n} \leq \frac{1}{c - c^2/L},$$

and it remains to let $L \rightarrow \infty$. \square

Remark 4.5. It might be tempting to deduce (16) from Proposition 4.2 directly using a suitable law of large numbers. However, a simple application of the Borel–Cantelli–Lévy lemma shows that we cannot do so without using the specifics of our stopping times τ_k . Indeed, assuming $c \in \{2, 3, \dots\}$, we can define a filtered probability space and stopping times τ_k , $k = 0, 1, \dots$, with $\tau_0 := 0$, in such a way that

$$\tau_k - \tau_{k-1} = \begin{cases} 1 & \text{with probability } 1 - k^{-2} \\ (c-1)k^2 + 1 & \text{with probability } k^{-2} \end{cases}$$

for all $k \in \mathbb{N}$ (where the probabilities may be conditional on a suitable σ -algebra $\mathcal{G}_{\tau_{k-1}}$). Then $\mathbb{E}(\tau_k - \tau_{k-1}) = c$ (and $\mathbb{E}(\tau_k - \tau_{k-1} \mid \mathcal{G}_{\tau_{k-1}}) = c$) for all k but, almost surely, $\tau_k - \tau_{k-1} = 1$ from some k on.

Of course, the statement of Proposition 4.4 also holds for the CUSUM procedure.

Corollary 4.6. *Let A_n be the number of alarms raised by the conformal CUSUM procedure (12) after seeing the observations z_1, \dots, z_n . Then (16) holds under the assumptions of Proposition 3.1.*

Proof. As in the proof of Corollary 4.3, combine Proposition 4.4 with the fact that Shiryaev–Roberts raises alarms more often than CUSUM does. \square

Optimality of procedures for change detection

It is remarkable that both CUSUM and Shiryaev–Roberts procedures are optimal under some natural conditions and for some natural criteria of optimality. As already mentioned, in standard settings of change detection the task is to detect a change from one known IID distribution for the incoming data to another known IID distribution. CUSUM and Shiryaev–Roberts procedures are then applied to a specific martingale, the likelihood ratio of the post-change distribution to the pre-change distribution. Therefore, they depend on just one parameter, the threshold c (for given pre-change and post-change distributions), whereas in the context of testing randomness we have a wide class of CUSUM and Shiryaev–Roberts procedures, built on top of different conformal martingales.

The five standard criteria for the quality of such specific procedures have been referred to by the letters A–E; see, e.g., Shiryaev [49]. Under natural conditions, Shiryaev–Roberts is optimal under two of the criteria, and CUSUM is optimal under one of them. Such statements of optimality are very satisfactory results about the efficiency of the corresponding procedures.

In this paper we only discuss validity results for CUSUM and Shiryaev–Roberts in the context of randomness testing and do not claim their optimality. As discussed at the end of Section 3, this is a difficult task already for the basic strongly valid testing procedure using conformal martingales. The null hypothesis (that of randomness) is composite and, moreover, very large (for large \mathbf{Z}), and we do not specify any alternatives; we simply do not have enough structure to specify a meaningful optimization problem.

5 IID probability vs exchangeability probability

We will be discussing two related interpretations of the efficiency of conformal martingales: on one hand, we can think of them as detecting deviations from randomness, and then “efficiency” would mean “ability to detect deviations from randomness in small samples”. On the other hand, we can think of them as detecting deviations from exchangeability, and the meaning of efficiency would change accordingly. In this section we discuss the relation between efficient testing for randomness and exchangeability in the simplest possible setting of finite binary data sequences and batch statistical tests. For infinite sequences randomness and exchangeability are connected by de Finetti’s theorem, as discussed in Remark 3.4; for finite sequences their relation becomes much less close. We re-connect our insights to conformal martingales in the next section.

In the 1960s Kolmogorov started revival of the interest in random sequences, believing that they are important for understanding the applications of probability theory and statistics. As already mentioned, he concentrated on binary sequences (as a simple starting point), in which context he often referred to them as *Bernoulli sequences*. His first imperfect publication on this topic was the 1963 paper [20] (Kolmogorov refers to it as “incomplete discussion”, ac-

ording to the English translation of [21]). In the same year he conceived using the notion of computability for formalizing randomness. Kolmogorov's main publications on the algorithmic theory of randomness were [21, 22, 23].

Let $\Omega := \{0, 1\}^N$ be the set of all binary sequences of a given length N , interpreted as sequences of observations. The time horizon $N \in \mathbb{N}$ can be regarded as fixed in the rest of this paper (apart from (27) and the appendix).

Let B_p be the Bernoulli probability measure on $\{0, 1\}$ with the probability of 1 equal to $p \in [0, 1]$: $B_p(\{1\}) := p$. The *upper IID probability* of a set $E \subseteq \Omega$ is defined to be

$$\mathbb{P}^{\text{iid}}(E) := \sup_{p \in [0, 1]} B_p^N(E), \quad (20)$$

and the *upper exchangeability probability* of $E \subseteq \Omega$ is defined to be

$$\mathbb{P}^{\text{exch}}(E) := \sup_P P(E), \quad (21)$$

P ranging over the exchangeable probability measures on Ω (in the current binary case we can say that a probability measure P on Ω is *exchangeable* if $P(\{\omega\})$ depends on $\omega \in \Omega$ only via the number of 1s in ω).

Remark 5.1. The *lower probabilities* corresponding to (20) and (21) are $1 - \mathbb{P}^{\text{iid}}(\Omega \setminus E)$ and $1 - \mathbb{P}^{\text{exch}}(\Omega \setminus E)$, respectively. In this paper we never need lower probabilities.

The function \mathbb{P}^{iid} can be used when testing the hypothesis of randomness: if $\mathbb{P}^{\text{iid}}(E)$ is small (say, below 1%) and the observed sequence ω is in E that is chosen in advance, we can reject the hypothesis that the observations in ω are IID. Similarly, \mathbb{P}^{exch} can be used when testing the hypothesis of exchangeability. This is an instance of application of *Cournot's principle*, often regarded to be the only bridge between probability theory and its applications. The principle was widely discussed at the beginning of the 20th century and defended by, e.g., Borel, Lévy, and Kolmogorov [43, Section 2.2]. Kolmogorov's statement of Cournot's principle in his *Grundbegriffe* [19, Chapter I, Section 2] is

If $P(A)$ is very small, then one can be practically certain that the event A will not occur on a single realization of the conditions \mathfrak{S} .

(The conditions \mathfrak{S} in this quote refer to the probability trial under discussion.) In the form stated by Kolmogorov, the principle goes back to Jacob Bernoulli [4] (see, e.g., [43, Section 2.2]). It establishes a bridge between probability theory and our expectations about reality; observing an event A (assumed to be chosen in advance) of a small probability casts doubt on P . Cournot's [7, p. 78] contribution was to state that this is the only bridge between probability theory and reality.

Cournot's principle suggests the following understanding of the efficiency of a method of testing the hypothesis of randomness: given any event E such that $\mathbb{P}^{\text{iid}}(E)$ is very small, the method should allow us to reject the hypothesis of randomness after observing E .

From the point of view of Bayesian statistics, a possible justification of Cournot's principle is the possibility of transforming p-values into Bayes factors; e.g., the function $p \mapsto 2\sqrt{p}$ transforms p-values into Bayes factors in favor of the null hypothesis (small values serving as evidence against the null hypothesis). For details, see, e.g., [42, Section 6]. Therefore, belonging to an event A of a small probability transforms into a small p-value, which in turn transforms into a small Bayes factor in favor of the null hypothesis.

Remark 5.2. The reciprocal function $p \mapsto p^{-1/2}/2$ transforms p-values into Bayes factors against the null hypothesis (with large values serving as evidence against the null hypothesis). This function belongs to the parametric family (9), all of which satisfy this property [42, Section 6]. For simple null hypotheses Bayes factors against the null hypothesis are the same thing as e-values [61, 13, 41], already mentioned in Section 3, but the difference becomes important for composite null hypotheses.

A popular way of getting rid of the parameter κ , popularized by Berger and his co-authors (see, e.g., [3]) is to minimize over it, getting

$$\max_{\kappa \in (0,1)} \kappa p^{\kappa-1} / = -\frac{1}{ep \log p}.$$

This value is sometimes (e.g., in the JASP package) referred to as the Vovk-Sellke maximum p-ratio (it was introduced independently in [57] and [40]). It is not useful for us in this paper since it is not a valid Bayes factor.

Proposition 5.3. *For any $E \subseteq \Omega$,*

$$\mathbb{P}^{\text{iid}}(E) \leq \mathbb{P}^{\text{exch}}(E) \leq 1.5\sqrt{N} \mathbb{P}^{\text{iid}}(E). \quad (22)$$

Proof. The first inequality in (22) follows from each IID measure on Ω being exchangeable. If E contains either the all-0 sequence $0 \dots 0$ or the all-1 sequence $1 \dots 1$, the second inequality in (22) is obvious ($\mathbb{P}^{\text{iid}}(E) = \mathbb{P}^{\text{exch}}(E) = 1$). If E is empty, it is also obvious ($\mathbb{P}^{\text{iid}}(E) = \mathbb{P}^{\text{exch}}(E) = 0$). Finally, if E is nonempty and contains neither sequence, we have, for some $k \in \{1, \dots, N-1\}$,

$$\mathbb{P}^{\text{exch}}(E) = \mathbb{P}^{\text{exch}}(E \cap \Omega_k) = \frac{1/\binom{N}{k}}{(k/N)^k (1-k/N)^{N-k}} \mathbb{P}^{\text{iid}}(E \cap \Omega_k) \quad (23)$$

$$\leq \frac{k!(N-k)!N^N}{N!k^k(N-k)^{N-k}} \mathbb{P}^{\text{iid}}(E) \leq \sqrt{2\pi}e^{1/6} \sqrt{\frac{k(N-k)}{N}} \mathbb{P}^{\text{iid}}(E) \quad (24)$$

$$\leq (\sqrt{2\pi}e^{1/6}/2)\sqrt{N} \mathbb{P}^{\text{iid}}(E) \leq 1.5\sqrt{N} \mathbb{P}^{\text{iid}}(E), \quad (25)$$

where Ω_k is the set of all sequences in Ω containing k 1s. The first equality in (23) follows from each exchangeable probability measure on Ω being a convex mixture of the uniform probability measures on Ω_k , $k = 0, \dots, N$. The second equality in (23) follows from the maximum of $B_p(\{\omega\})$, $\omega \in \Omega_k$, over $p \in [0, 1]$ being attained at $p = k/N$. The first inequality in (24) is equivalent to the

obvious $\mathbb{P}^{\text{iid}}(E \cap \Omega_k) \leq \mathbb{P}^{\text{iid}}(E)$. The second inequality in (24) follows from Stirling's formula

$$n! = \sqrt{2\pi n} n^{n+1/2} e^{-n} e^{r_n}, \quad 0 < r_n < \frac{1}{12n}, \quad (26)$$

valid for all $n \in \mathbb{N}$; see, e.g., [37], where it is also shown that $r_n > \frac{1}{12n+1}$. The first inequality in (25) follows from $\max_{p \in [0,1]} p(1-p) = 1/4$. \square

Remark 5.4. The constant 1.5 in inequality (22) is not too far from being optimal: when $N = 2$ and $E = \{(0, 1)\}$, it can be improved only to $\sqrt{2} \approx 1.414$. Notice that our argument in fact gives $\sqrt{2\pi}e^{1/6}/2 \approx 1.481$ instead of 1.5.

Kolmogorov's [22, 23] implicit interpretation of (22) was that \mathbb{P}^{iid} and \mathbb{P}^{exch} are close; on the log scale we have

$$-\log \mathbb{P}^{\text{iid}}(E) = -\log \mathbb{P}^{\text{exch}}(E) + O(\log N), \quad (27)$$

whereas typical values of $-\log \mathbb{P}^{\text{iid}}(E)$ and $-\log \mathbb{P}^{\text{exch}}(E)$ have the order of magnitude N for small (but non-zero) $|E|$. See Appendix A for further details.

From the point of view of Cournot's principle, Proposition 5.3 may be interpreted as saying that there is not much difference between testing randomness and testing exchangeability. If we have a test with critical region E of size ϵ for testing exchangeability, we can use it for testing randomness and its size will not increase; in the opposite direction, if we have a test with critical region E of size ϵ for testing randomness, we can use it for testing exchangeability, and its size will increase to at most $1.5\sqrt{N}\epsilon$. On the log-scale of Equation (27) the difference between the evidence provided by E against the hypothesis of randomness and against the hypothesis of exchangeability is $O(\log N)$; it is clear that the left-hand side of (27) can be as large as N (for a non-empty E and assuming that the logarithms are binary). In the algorithmic theory of randomness it is customary to ignore such differences, although from the point of view of statistics, the difference is substantial.

6 Conformal probability

In this section we explore the efficiency of conformal martingales, restricting ourselves to the simple case of a finite horizon N (as in the previous section). First we will define upper conformal probability \mathbb{P}^{conf} , an analogue of \mathbb{P}^{iid} and \mathbb{P}^{exch} for testing randomness using conformal martingales. Our simple version of upper conformal probability will be sufficient for our current purpose; there are other natural definitions. The *upper conformal probability* of $E \subseteq \Omega$ is

$$\mathbb{P}^{\text{conf}}(E) := \inf\{S_0 : \forall (z_1, \dots, z_N) \in E : S_N(z_1, \theta_1, z_2, \theta_2, \dots) \geq 1 \text{ } \theta\text{-a.s.}\}, \quad (28)$$

where S ranges over the nonnegative conformal martingales, " θ -a.s." refers to the uniform probability measure over $(\theta_1, \theta_2, \dots) \in [0, 1]^\infty$, and S_0 stands for

the constant $S_0(z_1, \theta_1, z_2, \theta_2, \dots)$. The definition (28) is in the spirit of [44, Section 2.1]; $\mathbb{P}^{\text{conf}}(E) < \epsilon$ for a small $\epsilon > 0$ means that there exists a nonnegative conformal martingale with a small initial value, below ϵ , that almost surely increases its value manyfold, to at least 1, if the event E happens. Therefore, we do not expect this event to happen under the hypothesis of randomness. This is spelled out in the following lemma.

Lemma 6.1. *For any event E , $\mathbb{P}^{\text{iid}}(E) \leq \mathbb{P}^{\text{conf}}(E)$.*

Proof. Let P be an IID measure on Ω and S be a nonnegative conformal martingale satisfying the condition in (28). It suffices to prove

$$P(E) \leq S_0; \tag{29}$$

indeed, we can then obtain $\mathbb{P}^{\text{iid}}(E) \leq \mathbb{P}^{\text{conf}}(E)$ by taking sup of the left-hand side of (29) over P and taking inf of the right-hand side over S .

To check (29), remember that $S_N \geq 1_E$ a.s., where 1_E is the indicator of E . Since S is a nonnegative martingale under P (like any nonnegative conformal martingale), we have

$$P(E) = \int 1_E dP \leq \int S_N dP = S_0. \quad \square$$

We will use upper conformal probability to make the notion of efficiency for conformal martingales more precise. Namely, if \mathbb{P}^{iid} and \mathbb{P}^{conf} are shown to be close, this could be interpreted as conformal martingales being able to detect any deviations from randomness. By Cournot's principle, any deviations from randomness are demonstrated by indicating in advance an event E of small probability under any IID measure, i.e., such that $\mathbb{P}^{\text{iid}}(E)$ is small, which then happens. If \mathbb{P}^{iid} and \mathbb{P}^{conf} are close, $\mathbb{P}^{\text{conf}}(E)$ will also be small. This means that there exists a nonnegative conformal martingale S that increases its value manyfold when E happens. We can choose S in advance since E is chosen in advance. And such an S will be successful whenever E is.

The following proposition shows that \mathbb{P}^{exch} and \mathbb{P}^{conf} are close, in the sense similar to the closeness of \mathbb{P}^{iid} and \mathbb{P}^{exch} asserted in Proposition 5.3 (see also (27)).

Proposition 6.2. *For any $E \subseteq \Omega$,*

$$\mathbb{P}^{\text{exch}}(E) \leq \mathbb{P}^{\text{conf}}(E) \leq N \mathbb{P}^{\text{exch}}(E). \tag{30}$$

Proposition 6.2 is a statement of efficiency for conformal martingales. It says that, at our crude scale, lack of exchangeability can be detected using conformal martingales. Namely, given a critical region E of a very small size $\epsilon := \mathbb{P}^{\text{exch}}(E)$, we can construct a nonnegative conformal martingale with initial capital $N\epsilon$ or less that attains capital of at least 1 when E happens.

Combining the right-hand sides of (22) and (30) we obtain

$$\mathbb{P}^{\text{conf}}(E) \leq 1.5N^{1.5} \mathbb{P}^{\text{iid}}(E).$$

This inequality says that conformal martingales are efficient at detecting deviations not only from exchangeability but also from randomness. Given a critical region E of a very small size $\epsilon := \mathbb{P}^{\text{iid}}(E)$, there exists a nonnegative conformal martingale that increases an initial capital of $1.5N^{1.5}\epsilon$ to at least 1 when E happens.

Proof of Proposition 6.2. First we check the left inequality in (30). We will do even more: we will check that it remains true even if the right-hand side of (28) is replaced by

$$\inf\{S_0 : \forall(z_1, \dots, z_N) \in E : \mathbb{E}_\theta S_N(z_1, \theta_1, z_2, \theta_2, \dots) \geq 1\}, \quad (31)$$

where the \mathbb{E}_θ refers to the uniform probability measure over $(\theta_1, \theta_2, \dots) \in [0, 1]^\infty$. Notice that S_0, \dots, S_N in (28) is a martingale in the filtration (\mathcal{G}_n) generated by the p-values p_1, \dots, p_N under any exchangeable probability measure on Ω ; this follows from the fact that p_1, \dots, p_N are IID and uniform on $[0, 1]$ under any exchangeable probability measure (see Remark 3.2). Therefore, for each $E \subseteq \Omega$ and each nonnegative conformal martingale S such that $\mathbb{E}_\theta S_N \geq 1_E$, we have

$$\mathbb{P}_z(E) \leq \mathbb{P}_z(\mathbb{E}_\theta S_N \geq 1) \leq \mathbb{E}_z(\mathbb{E}_\theta S_N) = \mathbb{E}_{z,\theta} S_N = S_0, \quad (32)$$

where \mathbb{P}_z refers to $(z_1, \dots, z_N) \sim P$, P is an exchangeable probability measure on Ω , \mathbb{E}_θ refers to $(\theta_1, \dots, \theta_N) \sim U^N$, U is the uniform probability measure on $[0, 1]$, and $\mathbb{E}_{z,\theta}$ refers to $(z_1, \dots, z_N) \sim P$ and $(\theta_1, \dots, \theta_N) \sim U^N$ independently. Taking the sup of the leftmost expression in (32) over P and the inf of the rightmost expression in (32) over S , we obtain the left inequality in (30).

It remains to check the right inequality in (30). Let us first check the part “ \leq ” of the first equality in

$$\mathbb{P}^{\text{conf}}(\{\omega\}) = \frac{k!(N-k)!}{N!} = \mathbb{P}^{\text{exch}}(\{\omega\}),$$

where $k \in \{0, \dots, N\}$ and $\omega \in \Omega$ contains k 1s (the part “ \geq ” was established in the previous paragraph; it will not be used in the rest of this proof).

Let $\omega = (z_1, \dots, z_N)$ be the representation of ω as a sequence of bits. Consider the nonnegative conformal martingale S^ω obtained from the identity non-conformity measure $A(z_1, \dots, z_n) := (z_1, \dots, z_n)$ and a betting martingale F such that $F(\square) = 1/\binom{N}{k}$ (where \square is the empty sequence) and

$$\frac{F(p_1, \dots, p_{n-1}, p_n)}{F(p_1, \dots, p_{n-1})} := \begin{cases} \frac{n}{k_n} & \text{if } p_n \leq k_n/n \text{ and } z_n = 1 \\ \frac{n}{n-k_n} & \text{if } p_n \geq k_n/n \text{ and } z_n = 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $n = 1, \dots, N$ and k_n is the number of 1s in ω observed so far,

$$k_n := |\{j \in \{1, \dots, n\} : z_j = 1\}|;$$

in particular, $k_N = k$. (Intuitively, S^ω gambles recklessly on the n th observation being z_n .) If the actual sequence of observations happens to be ω , on step n the value of the martingale S^ω is multiplied, a.s., by the fraction whose numerator is n and whose denominator is the number of bits z_n observed in ω so far. The product of all these fractions over $n = 1, \dots, N$ will have $N!$ as its numerator and $k!(N-k)!$ as its denominator. This conformal martingale is almost deterministic, in the sense of not depending on θ_n provided $\theta_n \notin \{0, 1\}$, and its final value on ω is, a.s.,

$$\frac{1}{\binom{N}{k}} \frac{N!}{k!(N-k)!} = 1.$$

To move from singletons to arbitrary $E \subseteq \Omega$, notice that a finite linear combination of conformal martingales S^ω with positive coefficients is again a conformal martingale, since they involve the same nonconformity measure, and betting martingales can be combined. Fix $E \subseteq \Omega$ and remember that Ω_k is the set of all sequences in Ω containing k 1s. Represent E as the disjoint union

$$E = \bigcup_{k=0}^N E_k, \quad E_k \subseteq \Omega_k,$$

and let U_k be the uniform probability measure on Ω_k . We then have

$$\begin{aligned} \mathbb{P}^{\text{conf}}(E) &\leq \sum_{\omega \in E} \mathbb{P}^{\text{conf}}(\{\omega\}) = \sum_{k=0}^N \sum_{\omega \in E_k} \mathbb{P}^{\text{conf}}(\{\omega\}) = \sum_{k=0}^N \sum_{\omega \in E_k} \mathbb{P}^{\text{exch}}(\{\omega\}) \\ &= \sum_{k=0}^N U_k(E_k) \leq N \max_{k=0, \dots, N} U_k(E_k) = N \mathbb{P}^{\text{exch}}(E), \end{aligned}$$

where the last inequality holds when, e.g., E does not contain the all-0 sequence $0 \dots 0 \in \Omega$. If E does contain the all-0 sequence, it is still true that

$$\mathbb{P}^{\text{conf}}(E) \leq 1 \leq N = N \mathbb{P}^{\text{exch}}(E). \quad \square$$

7 Conclusion

This paper gives a review of known methods of testing randomness online, all of which are based on conformal martingales. It raises plenty of questions, without giving many answers. Propositions 5.3 and 6.2 say that IID, exchangeability, and conformal upper probabilities are close, but the accuracy of these statements is very low and far from meaningful in practice. The most obvious direction of further research is to obtain more accurate results (a simple example related to Proposition 5.3 will be given in Appendix A). It would be ideal to establish exact bounds on upper conformal probability in terms of upper IID probability and upper exchangeability probability. The most natural definition of upper conformal probability in this context might involve randomness in a more substantial way than our official definition (28) does (cf., e.g., (31)).

Acknowledgements

The original version of this paper was written in support of my poster at ISIPTA 2019 (Eleventh International Symposium on Imprecise Probabilities: Theories and Applications) presented on 5 July 2019. I am grateful to all discussants, including Thomas Dietterich, Wouter Koolen, and Glenn Shafer. This research has been supported by Amazon, Astra Zeneca, and Stena Line. The experiments with the tangent metric are based on Daniel Keyzers’s [18] implementation in C. Bartels’s test for randomness used in Section 2 is from the R package `randtests` (Testing randomness in R) by Frederico Caeiro and Ayana Mateus [6]. Comments by three reviewers have been extremely useful for improving the presentation; they led, in particular, to the inclusion of experiments with the `Absenteeism` dataset and to the explicit discussion of Cournot’s principle.

References

- [1] Vineeth N. Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk, editors. *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations, and Applications*. Elsevier, Amsterdam, 2014.
- [2] Robert Bartels. The rank version of von Neumann’s ratio test for randomness. *Journal of the American Statistical Association*, 77:40–46, 1982.
- [3] Daniel J. Benjamin and James O. Berger. Three recommendations for improving the use of p-values. *American Statistician*, 73(Supplement 1):186–191, 2019.
- [4] Jacob Bernoulli. *Ars Conjectandi*. Thurnisius, Basel, 1713.
- [5] P. K. Bhattacharya. Some aspects of change-point analysis. In *Change-point Problems*, volume 23 of *IMS Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, Hayward, CA, 1994.
- [6] Frederico Caeiro and Ayana Mateus. An R implementation of several randomness tests. *AIP Conference Proceedings*, 1618:531–534, 2014.
- [7] Antoine-Augustin Cournot. *Exposition de la théorie des chances et des probabilités*. Hachette, Paris, 1843.
- [8] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, Hoboken, NJ, second edition, 2006.
- [9] Joseph L. Doob. *Stochastic Processes*. Wiley, New York, 1953.
- [10] Wenyu Du, Aleksey S. Polunchenko, and Grigory Sokolov. On robustness of the Shiryaev–Roberts change-point detection procedure under parameter misspecification in the post-change distribution. *Communications in Statistics—Simulation and Computation*, 46:2185–2206, 2017.

- [11] Dheeru Dua and Casey Graff. UCI machine learning repository, 2020. The **Absenteeism at work** dataset was donated by Andrea Martiniano, Ricardo Pinto Ferreira, and Renato Jose Sassi.
- [12] Valentina Fedorova, Alex Gammerman, Ilia Nouretdinov, and Vladimir Vovk. Plug-in martingales for testing exchangeability on-line, On-line Compression Modelling project (New Series), <http://alrw.net>, Working Paper 4, April 2012. Conference version: ICML 2012.
- [13] Peter Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing. Technical Report arXiv:1906.07801 [math.ST], arXiv.org e-Print archive, June 2020.
- [14] Mark Herbster and Manfred K. Warmuth. Tracking the best expert. *Machine Learning*, 32:151–178, 1998.
- [15] Shen-Shyang Ho and Harry Wechsler. A martingale framework for detecting changes in data streams by testing exchangeability. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:2113–2127, 2010.
- [16] Harold Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, third edition, 1961.
- [17] John L. Kelly. A new interpretation of information rate. *Bell System Technical Journal*, 35:917–926, 1956.
- [18] Daniel M. Keyzers. Approaches to invariant image object recognition. Master’s thesis, Lehrstuhl für Informatik VI, Mathematisch-Naturwissenschaftliche Fakultät, Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany, June 2000.
- [19] Andrei N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933. English translation: *Foundations of the Theory of Probability*. Chelsea, New York, 1950.
- [20] Andrei N. Kolmogorov. On tables of random numbers. *Sankhya. Indian Journal of Statistics A*, 25:369–376, 1963.
- [21] Andrei N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1:1–7, 1965.
- [22] Andrei N. Kolmogorov. Logical basis for information theory and probability theory. *IEEE Transactions on Information Theory*, IT-14:662–664, 1968.
- [23] Andrei N. Kolmogorov. Combinatorial foundations of information theory and the calculus of probabilities. *Russian Mathematical Surveys*, 38:29–40, 1983.

- [24] Andrei N. Kolmogorov. On logical foundations of probability theory. In Yu. V. Prokhorov and K. Itô, editors, *Probability Theory and Mathematical Statistics*, volume 1021 of *Lecture Notes in Mathematics*, pages 1–5. Springer, 1983. Talk at the Fourth USSR–Japan Symposium on Probability Theory and Mathematical Statistics (Tbilisi, August 1982) recorded by Alexander A. Novikov, Alexander K. Zvonkin, and Alexander Shen.
- [25] Andrei N. Kolmogorov and Albert N. Shiryaev. Применение марковских процессов к обнаружению разладок производственного процесса. Доклад на VI Совещании по теории вероятностей и математической статистике, Vilnius, 1960.
- [26] Andrei N. Kolmogorov and Vladimir A. Uspensky. Algorithms and randomness. *Theory of Probability and Its Applications*, 32:389–412, 1987.
- [27] Erich L. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco, 1975.
- [28] Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, New York, third edition, 2005.
- [29] Per Martin-Löf. The definition of random sequences. *Information and Control*, 9:602–619, 1966.
- [30] Per Martin-Löf. Personal communication, February 2005.
- [31] David McDonald. A cusum procedure based on sequential ranks. *Naval Research Logistics*, 37:627–646, 1990.
- [32] Richard von Mises. Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 5:52–99, 1919.
- [33] Ewan S. Page. Continuous inspection schemes. *Biometrika*, 41:100–115, 1954.
- [34] Moshe Pollak. Average run lengths of an optimal method of detecting a change in distribution. *Annals of Statistics*, 15:749–779, 1987.
- [35] H. Vincent Poor and Olympia Hadjiladis. *Quickest Detection*. Cambridge University Press, Cambridge, 2009.
- [36] Karl R. Popper. *The Open Universe*. Hutchinson, London, 1982.
- [37] Herbert Robbins. A remark on Stirling’s formula. *American Mathematical Monthly*, 62:26–29, 1955.
- [38] S. W. Roberts. A comparison of some control chart procedures. *Technometrics*, 8:411–430, 1966.
- [39] Mark J. Schervish. *Theory of Statistics*. Springer, New York, 1995.

- [40] Thomas Sellke, M. J. Bayarri, and James Berger. Calibration of p-values for testing precise null hypotheses. *American Statistician*, 55:62–71, 2001.
- [41] Glenn Shafer. The language of betting as a strategy for statistical and scientific communication. Technical Report arXiv:1903.06991 [math.ST], arXiv.org e-Print archive, March 2019. To appear as discussion paper in the *Journal of the Royal Statistical Society A*, and to be read in September 2020.
- [42] Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors, and p-values. *Statistical Science*, 26:84–101, 2011.
- [43] Glenn Shafer and Vladimir Vovk. The sources of Kolmogorov’s *Grundbegriffe*. *Statistical Science*, 21:70–98, 2006.
- [44] Glenn Shafer and Vladimir Vovk. *Game-Theoretic Foundations for Probability and Finance*. Wiley, Hoboken, NJ, 2019.
- [45] Walter A. Shewhart. *Economic Control of Quality of Manufactured Product*. Van Nostrand, New York, 1931.
- [46] Albert N. Shiryaev. On optimum methods in quickest detection problems. *Theory of Probability and Its Applications*, 8:22–46, 1963.
- [47] Albert N. Shiryaev. Quickest detection problems: Fifty years later. *Sequential Analysis*, 29:345–385, 2010. Editor’s special invited paper.
- [48] Albert N. Shiryaev. *Probability-2*. Springer, New York, third edition, 2019.
- [49] Albert N. Shiryaev. *Stochastic Disorder Problems*. Springer, Cham, Switzerland, 2019.
- [50] David Siegmund, Benjamin Yakir, and Nancy R. Zhang. Detecting simultaneous variant intervals in aligned sequences. *Annals of Applied Statistics*, 5:645–668, 2011.
- [51] Patrice Simard, Yann LeCun, and John Denker. Efficient pattern recognition using a new transformation distance. *Advances in Neural Information Processing Systems*, 5:50–58, 1993.
- [52] Vladimir A. Uspensky. Колмогоров, каким я его помню (Kolmogorov as I remember him). In Albert N. Shiryaev, editor, *Колмогоров в воспоминаниях (Kolmogorov in Remembrance)*. Moscow, Nauka, 1993.
- [53] Tim van Erven, Peter Grünwald, and Steven de Rooij. Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the AIC–BIC dilemma. *Journal of the Royal Statistical Society B*, 74:361–417, 2012.

- [54] Jean Ville. *Etude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.
- [55] Denis Volkhonskiy, Evgeny Burnaev, Ilia Nouretdinov, Alex Gammerman, and Vladimir Vovk. Inductive conformal martingales for change-point detection. *Proceedings of Machine Learning Research*, 60:132–153, 2017. COPA 2017.
- [56] Vladimir Vovk. On the concept of the Bernoulli property. *Russian Mathematical Surveys*, 41:247–248, 1986. Russian original: О понятии бернуллиевости. Another English translation with proofs: On-line Compression Modelling project (New Series), <http://alrw.net>, Working Paper 15.
- [57] Vladimir Vovk. A logic of probability, with application to the foundations of statistics (with discussion). *Journal of the Royal Statistical Society B*, 55:317–351, 1993.
- [58] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.
- [59] Vladimir Vovk and Vladimir V. V'yugin. On the empirical validity of the Bayesian method. *Journal of the Royal Statistical Society B*, 55:253–266, 1993.
- [60] Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 2020. To appear, published online.
- [61] Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination, and applications. Technical Report arXiv:1912.06116 [math.ST], arXiv.org e-Print archive, September 2020.
- [62] Abraham Wald. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, 16:117–186, 1945.
- [63] Abraham Wald. *Sequential Analysis*. Wiley, New York, 1947.
- [64] Abraham Wald and Jacob Wolfowitz. An exact test for randomness in the non-parametric case based on serial correlation. *Annals of Mathematical Statistics*, 14:378–388, 1943.
- [65] Abraham Wald and Jacob Wolfowitz. Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics*, 19:326–339, 1948.
- [66] Xiao Zhang, Peter Lu, Josée Martens, Gary Ericson, and Kent Sharkey. *Time Series Anomaly Detection module in Microsoft Azure*. Microsoft, Seattle, WA, May 2020. Online documentation.

A Connections with the algorithmic theory of randomness

The emphasis of this appendix will be on Kolmogorov's approach to randomness and exchangeability expressed in Martin-Löf's [29] terms, which are closer to the traditional statistical language. (Kolmogorov's original definitions, equivalent but given in terms of algorithmic complexity, is discussed in Appendix B.) In our terminology we will follow [59]. Following Kolmogorov [21, 22, 23], we will only consider the case of binary observations.

A *measure of randomness* is an upper semicomputable function $f : \{0, 1\}^* \rightarrow [0, 1]$ such that, for any $N \in \mathbb{N}$, any IID measure P on $\{0, 1\}^N$, and any $\epsilon > 0$, we have (3). The upper semicomputability of f means that there exists an algorithm that, when fed with a rational number r and sequence $\omega \in \{0, 1\}^*$, eventually stops if $f(\omega) < r$ and never stops otherwise.

In other words, a measure of randomness is a family of p-variables for testing randomness in $\{0, 1\}^N$. The requirement of upper semicomputability is natural: e.g., if $f(\omega) < 0.01$ (the p-value is highly statistically significant), we should learn this eventually.

Analogously, a *measure of exchangeability* is an upper semicomputable function $f : \{0, 1\}^* \rightarrow [0, 1]$ such that, for any $N \in \mathbb{N}$, any exchangeable measure P on $\{0, 1\}^N$, and any $\epsilon > 0$, we have (3).

Lemma A.1. *There exists a measure of randomness f (called universal) such that any other measure of randomness f' satisfies $f = O(f')$. There exists a measure of exchangeability f (called universal) such that any other measure of exchangeability f' satisfies $f = O(f')$.*

The proof of Lemma A.1 is standard; see, e.g., [29] or [59, Lemma 4].

In the algorithmic theory of randomness, it is customary to measure lack of randomness or exchangeability on the log scale. Therefore, we fix a universal measure of randomness f , set $d^{\text{iid}} := -\log f$, and refer to $d^{\text{iid}}(\omega)$ as the *deficiency of randomness* of the sequence $\omega \in \{0, 1\}^*$. Similarly, we fix a universal measure of exchangeability f , set $d^{\text{exch}} := -\log f$, and refer to $d^{\text{exch}}(\omega)$ as the *deficiency of exchangeability* of ω . (Traditionally, the log is binary.)

Proposition 5.3 immediately implies

$$d^{\text{exch}}(\omega) - O(1) \leq d^{\text{iid}}(\omega) \leq d^{\text{exch}}(\omega) + \frac{1}{2} \log N + O(1), \quad (33)$$

where ω ranges over $\{0, 1\}^*$ and N is the length of ω . In fact, we can interpret (33) as the algorithmic version of Proposition 5.3. Kolmogorov regarded the coincidence to within log as close enough, at least for some purposes: cf. the last two paragraphs of [22]; therefore, he preferred the simpler definition $d^{\text{exch}}(\omega) \approx 0$ of ω being a Bernoulli sequence.

Proposition 5.3 is very crude, and Section 7 sets the task of obtaining more accurate results. In fact, such results are known in the context of the algorithmic theory of randomness; some were obtained in the paper [56] written under Kolmogorov's supervision.

To clarify relations between algorithmic randomness and exchangeability, we will need another notion, binomiality. The binomial probability distribution $\text{bin}_{N,p}$ on $\{0, \dots, N\}$ with parameter p is defined by

$$\text{bin}_{N,p}(\{k\}) := \binom{N}{k} p^k (1-p)^{N-k}, \quad k \in \{0, \dots, N\}.$$

A *measure of binomiality* is an upper semicomputable function $f : \{(N, k) : N \in \mathbb{N}, k \in \{0, \dots, N\}\} \rightarrow [0, 1]$ such that, for any $N \in \mathbb{N}$, any $p \in [0, 1]$, and any $\epsilon > 0$,

$$\text{bin}_{N,p}(\{k : f(N, k) \leq \epsilon\}) \leq \epsilon.$$

Lemma A.2. *There exists a measure of binomiality f (called universal) such that any other measure of binomiality f' satisfies $f = O(f')$.*

We fix a universal measure of binomiality f , set $d^{\text{bin}}(k; N) := -\log f(N, k)$, and refer to $d^{\text{bin}}(k; N)$ as the *deficiency of binomiality* of k (in $\{0, \dots, N\}$).

Proposition A.3. *For any constant $\epsilon > 0$,*

$$\begin{aligned} (1 - \epsilon) (d^{\text{exch}}(\omega) + d^{\text{bin}}(k; N)) - O(1) &\leq d^{\text{iid}}(\omega) \\ &\leq (1 + \epsilon) (d^{\text{exch}}(\omega) + d^{\text{bin}}(k; N)) + O(1), \end{aligned}$$

N ranging over \mathbb{N} , ω over $\{0, 1\}^N$, and k being the number of 1s in ω .

Proposition A.3 follows immediately from (and is stated, in a more precise form, after) [56, Theorem 1]. It says, informally, that the randomness of ω is equivalent to the conjunction of two conditions: ω should be exchangeable, and the number of 1s in it should be binomial. For example, suppose that N is a large even number and the number of 1s in $\omega \in \{0, 1\}^N$ is $k = N/2$. Then ω might be perfectly exchangeable whereas it will not be random since it belongs to the set of all binary sequences with the number of 1s precisely $N/2$, whose probability (4) is small. (Example 2.2 was based on this observation expressed in a different language.)

B Connections with the algorithmic theory of complexity

In Appendix A we gave the definition of randomness equivalent to Kolmogorov's but used Martin-Löf's language of statistical tests. Kolmogorov himself used the language of algorithmic complexity, nowadays known as Kolmogorov complexity. Apart from his papers [21, 22, 23] on this topic, Kolmogorov was also a co-author of [26], which was based on his ideas and publications, although he did not see the final version of that paper [26, Introduction] and did not take part in preparing the talk on which it was based [52, p. 380]. Another valuable source is the record of his 1982 talk [24]. From now on I will assume knowledge of some basic notions of the theory of Kolmogorov complexity.

Kolmogorov's original notion of randomness for an element ω of a simple finite set M was that $K(\omega) \approx -\log |M|$, where K is Kolmogorov complexity and \log is binary log (see [21, Section 4]). Martin-Löf [30] modified this requirement to $K(\omega \mid M) \approx -\log |M|$, where K now stands for conditional Kolmogorov complexity. In his 1968 paper [22, Section 2] Kolmogorov gave his alternative formalization of von Mises's random sequences, with a reference to Martin-Löf: namely, Kolmogorov said that a binary sequence ω of length N containing k 1s is *Bernoulli* if

$$K(\omega \mid k, N) \approx \log \binom{N}{k}.$$

It is natural to call the difference

$$d^{\text{exch}}(\omega) := \log \binom{N}{k} - K(\omega \mid k, N) \quad (34)$$

the *deficiency of exchangeability* of ω (in terminology close to that of [24] and [26]). This definition is equivalent to the one given in Appendix A (in the sense that the difference between (34) and $d^{\text{exch}}(\omega)$ as defined in Appendix A is bounded in absolute value by a constant independent of ω ; this is the first italicized statement in [29, p. 616]).

Being Bernoulli in the sense of Kolmogorov does not fully reflect the intuition of being random, i.e., being a plausible outcome of a sequence of N tosses of a possibly biased coin. This intuition is better captured by

$$d^{\text{iid}}(\omega) := \inf_{p \in [0,1]} (-\log B_p^N(\omega) - K(\omega \mid p, N)), \quad (35)$$

which we call the *deficiency of randomness* of ω , being small. This is equivalent to the definition given in Appendix A.

Proposition A.3 in Appendix A about the difference between (34) and (35) can be made much more precise if we modify our definitions. Theorem 1 of [56] shows that

$$D^{\text{iid}}(\omega) = \left(\log \binom{N}{k} - KP(\omega \mid N, k, D^{\text{bin}}(k; N)) \right) + D^{\text{bin}}(k; N) + O(1),$$

where N ranges over \mathbb{N} , ω over $\{0, 1\}^N$, k is the number of 1s in ω , KP is prefix complexity, D^{iid} is the analogue of d^{iid} using prefix instead of Kolmogorov complexity, and $D^{\text{bin}}(k; N)$ is the *prefix deficiency of binomiality* of k defined by

$$D^{\text{bin}}(k; N) := \inf_{p \in [0,1]} (-\log \text{bin}_{N,p}(k) - KP(k \mid p, N)),$$

where $\text{bin}_{N,p}$ is the binomial probability measure on $\{0, \dots, N\}$ with parameter p , as defined earlier. Theorem 2 of [56] characterizes $D^{\text{bin}}(k; N)$ in terms of prefix complexity, showing that it can be as large as $\frac{1}{2} \log N + O(1)$.