Proceedings of Machine Learning Research 152:1–20, 2021 Conformal and Probabilistic Prediction and Applications

Conformal testing in a binary model situation

Vladimir Vovk v.vovk@rhul.ac.uk
Centre for Reliable Machine Learning, Royal Holloway, University of London, Egham, Surrey, UK

Editor: Lars Carlsson, Zhiyuan Luo, Giovanni Cherubin, and Khuong An Nguyen

Abstract

Conformal testing is a way of testing the IID assumption based on conformal prediction. The topic of this paper is experimental evaluation of the performance of conformal testing in a model situation in which IID binary observations generated from a Bernoulli distribution are followed by IID binary observations generated from another Bernoulli distribution, with the parameters of the distributions and changepoint known or unknown. Existing conformal test martingales can be used for this task and work well in simple cases, but their efficiency can be improved greatly.

Keywords: conformal test martingales, exchangeability martingales, Bernoulli model, alternative hypothesis

1. Introduction

The method of conformal prediction can be adapted to testing the IID model (Vovk et al., 2005, Section 7.1). The usual testing procedures in mathematical statistics (Lehmann and Romano, 2005) are performed in the batch mode: we are looking for evidence against the null hypothesis when given a batch of data (a dataset of observations). Conformal testing is different in that it processes the observations sequentially (online), and the amount of evidence found against the null hypothesis is updated when new observations arrive. Online hypothesis testing, for various null hypotheses, has been promoted in, e.g., Shafer and Vovk (2019), Shafer (2021), Grünwald et al. (2020), and Ramdas et al. (2021). In this setting, valid testing procedures are equated with *test martingales*, i.e., nonnegative processes with initial value 1 that are martingales under the null hypothesis.

At this time conformal testing is the only known general online procedure for testing the IID model. Namely, conformal test martingales are the only known non-trivial examples of exchangeability martingales, i.e., online testing procedures valid under the IID assumption. An important application of such procedures is in deciding when to retrain an algorithm of machine learning; for details, see Vovk et al. (2021). This paper does not deal directly with such important applications and, instead, lays foundations for more efficient methods for making such decisions.

For a long time it had remained unclear how efficient conformal testing is, but Vovk (2021, Section 6) argues that in the binary case conformal testing is efficient at least in a crude sense. This paper confirms that claim using simulation studies in a simple model situation. More generally, it proposes a programme of research into the efficiency of conformal testing in various model situations. The idea is very standard (Neyman and Pearson, 1933): to complement the null hypothesis (namely, the IID model) by a specific alternative hypothesis and investigate the power of our methods (namely, conformal testing) under the alternative. Unlike the Neyman–Pearson setting, this will not lead to a well-defined optimization problem, but it will give us an informal goal, and we will still be able to design efficient "custom-made" test martingales.

An important by-product of the proposed programme is developing useful tricks for conformal testing that might be useful in applications. We will see examples in Section 5.

Our simulation studies will explore the performance of various test martingales, including conformal test martingales, and related processes, to be defined in Section 4. Conformal prediction uses randomization for tie-breaking, and this feature is inherited by conformal testing. In particular, conformal test martingales are randomized. All plots in this paper have been produced using the seed 2021 for the NumPy pseudorandom number generator, and the dependence on the seed does not change any of our conclusions.

Remark 1 In this paper we will avoid the expression "conformal martingale", as used in Vovk (2021), in order to avoid terminology clash with the notion of conformal martingale introduced in Getoor and Sharpe (1972) and discussed in Walsh (1977). (Even though this would not have led to any confusion; in general, the two notions are so different that they are unlikely to be used in the same context.)

2. Model situation

This section introduces the main model situation considered in this paper. Our data consist of binary observations generated independently from Bernoulli distributions. Let $B(\pi)$ be the Bernoulli distribution on $\{0, 1\}$ with parameter $\pi \in [0, 1]$: $B(\pi)(\{1\}) = \pi$. We assume that the observations are IID except that at some point the value of the parameter π changes. Let π_0 be the pre-change parameter and π_1 be the post-change parameter. The total number of observations is N, of which the first N_0 come from the pre-change distribution $B(\pi_0)$ and the remaining $N_1 := N - N_0$ from the post-change distribution $B(\pi_1)$.

Our main model situation is the one considered by Ramdas et al. (2021, Section 4). In their setting, $\pi_0 = 0.1$, $\pi_1 = 0.4$, $N = 10^4$, and $N_0 = N_1 = 5000$. Ramdas et al. construct

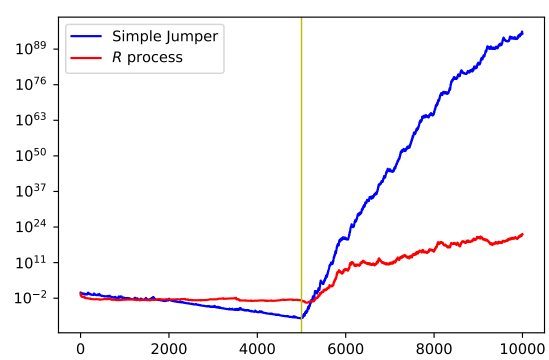

Figure 1: The process R of Ramdas et al. (2021) and the Simple Jumper martingale of Vovk et al. (2021), as described in text (neither designed for the changepoint detection problem).

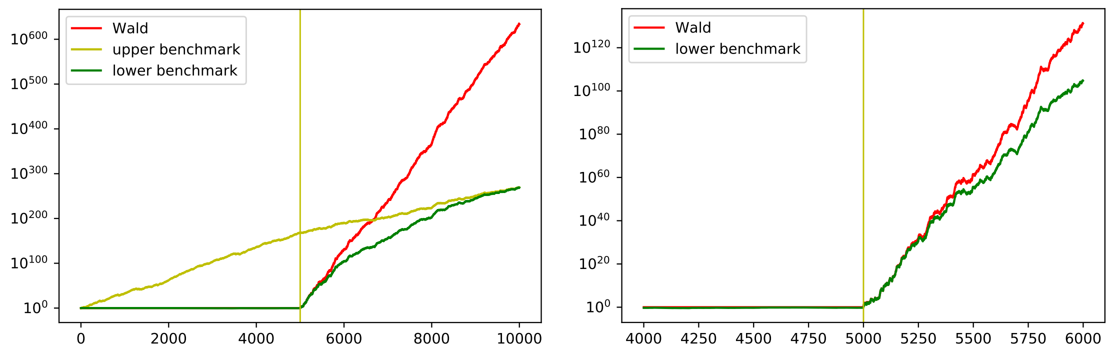

Figure 2: Left panel: Wald's martingale (red line), the upper benchmark (yellow line), and the lower benchmark (green line) over the whole dataset. Right panel (close-up of the left panel): Wald's martingale and the lower benchmark over the middle 2000 observations.

a process $R = R_n$ which, for any IID probability measure $B(\pi)^\infty$, is dominated by a test martingale $M_n^{(\pi)}$ w.r. to $B(\pi)^\infty$: $R_n \le M_n^{(\pi)}$ for all n and π. The trajectory of their process in the model situation is shown in Figure 1 in red (it coincides with the trajectory in Figure 3 in Ramdas et al. 2021 apart from using a different randomly generated dataset). Figure 1 shows in blue the trajectory of the Simple Jumper conformal test martingale, as defined in Vovk et al. (2021), based on the identity nonconformity measure; the martingale (including the parameter $J = 0.01$) is exactly as described in Vovk et al. (2021, Algorithm 1). Both processes can serve as measures of the amount of evidence found against the null hypothesis, and both perform very well finding decisive evidence against the null hypothesis.

Neither the process R nor the Simple Jumper martingale were designed for the changepoint detection problem. The process R was designed for the alternative being a Markov chain, and its good performance in the problem of changepoint detection was an interesting byproduct. The Simple Jumper martingale was designed in Vovk (2020c) to achieve a reasonable performance on the USPS dataset, without a clear alternative in mind. In this paper we will take the problem of changepoint detection more seriously. Our goal will be to explore attainable final values of test martingales in model situations such as that in Ramdas et al. (2021, Section 4) (our alternative hypotheses). Our null hypothesis is the *IID model*, under which the observations are IID but the value of the parameter π is unrestricted.

3. Two benchmarks

In this section we will discuss possible benchmarks that we can use for evaluating the quality of our conformal test martingales. For each $n \in \{1, 2, \dots\}$, let $k(n)$ be the number of 1s among the first n observations in the binary (consisting of 0 and 1) data sequence. In Sections 3–5 we consider our main model situation: $\pi_0 = 0.1$, $\pi_1 = 0.4$, $N = 10^4$, and $N_0 = N_1 = 5000$.

The first process that we discuss is the likelihood ratio of the true distribution to the pre-change distribution:

$$W_n := \begin{cases} 1 & \text{if } n \le N_0 \\ \left(\frac{\pi_1}{\pi_0}\right)^{k(n)-k(N_0)} \left(\frac{1-\pi_1}{1-\pi_0}\right)^{(n-N_0)-(k(n)-k(N_0))} & \text{otherwise.} \end{cases}$$

This is the optimal test martingale in Wald's (Wald, 1947; Wald and Wolfowitz, 1948) sense, and we will call it *Wald's martingale*. This process, however, is a test martingale only with respect to the null hypothesis $B(\pi_0)^\infty = B(0.1)^\infty$, whereas our null hypothesis is the IID model. Therefore, it is not a reasonable benchmark. Its trajectory is shown in red in Figure 2 (over the full dataset on the left, and over its middle part on the right).

Figure 2 shows in green the infimum of the likelihood ratios

$$L_n := \begin{cases} \frac{\pi_0^{k(n)}(1-\pi_0)^{n-k(n)}}{\left(\frac{k(n)}{n}\right)^{k(n)}\left(1-\frac{k(n)}{n}\right)^{n-k(n)}} & \text{if } n \le N_0 \\ \frac{\pi_0^{k(N_0)}(1-\pi_0)^{N_0-k(N_0)}\pi_1^{k(n)-k(N_0)}(1-\pi_1)^{(n-N_0)-(k(n)-k(N_0))}}{\left(\frac{k(n)}{n}\right)^{k(n)}\left(1-\frac{k(n)}{n}\right)^{n-k(n)}} & \text{otherwise} \end{cases} \tag{1}$$

(where $0^0 := 1$) of the true data distribution to $B(\pi)^\infty$ over π. We will refer to this process as the *lower benchmark*; its final value $\mathrm{LB}_N := L_N$ is indicative of the best result that can be attained in our testing problem.

Remark 2 The expression (1) is the infimum over the IID measures of the likelihood ratios that are individually optimal (for each IID measure) in Wald's sense. However, this does not mean that the infimum (1) itself is optimal. The extreme case for binary observations is where the null hypothesis consists of all probability measures on $\{0,1\}^\infty$. The analogue of the lower benchmark will quickly tend to 0, and so its performance will be much worse than that of the identical 1 (which is a test martingale under any null hypothesis). For more general observation spaces, such as in the case of real numbers changing their distribution (e.g., with $N(0,1)$ as pre-change distribution and $N(1,1)$ as post-change distribution), the IID model becomes too large, and we are in a situation that is even worse: the analogues of the ratios in (1) become zero. (Remember that such analogues have the supremum over all IID measures in the denominator, not the supremum over some parametric model containing both pre-change and post-change distributions.) The case of (1), however, is very far from these difficult cases, and even to the left of N_0 the trajectory of L_n is visually indistinguishable from 1.

Figure 2 shows that Wald's likelihood ratio process grows exponentially fast after the changepoint, which shows as a linear growth on the log scale. Its trajectory looks like a tangent to the lower benchmark trajectory. It is clear that the lower benchmark cannot grow exponentially fast: the post-change distribution $B(0.4)$ is gradually becoming "the new normal".

In order to develop an alternative to (1) that would also work outside the binary case, let us replace the denominator of (1), which is the maximum likelihood chosen *a posteriori*, by the likelihood at a parameter value chosen *a priori* but with the knowledge of the stochastic mechanism generating the data. Let us generalize our setting slightly, assuming that the observations take values in a finite set and take value i with probability $\pi_{0,i}$ before the changepoint and $\pi_{1,i}$ after the changepoint (so that $\sum_i \pi_{0,i} = \sum_i \pi_{1,i} = 1$). Our goal is to find a probability measure (u_i) for one observation such that the (random) likelihood ratio of the true data-generating distribution to the Nth power of (u_i) is as small as possible. By the Kelly criterion, the corresponding optimization problem for the optimal probability measure (u_i) in the denominator is

$$N_0 \sum_i \pi_{0,i} \ln \frac{\pi_{0,i}}{u_i} + N_1 \sum_i \pi_{1,i} \ln \frac{\pi_{1,i}}{u_i} \to \min,$$

which simplifies to

$$\sum_i \frac{N_0 \pi_{0,i} + N_1 \pi_{1,i}}{N} \ln u_i \to \max .$$

By the nonnegativity of Kullback–Leibler divergence, the optimal solution is

$$u_i := \frac{N_0 \pi_{0,i} + N_1 \pi_{1,i}}{N},$$

i.e., the weighted average of π_0 and π_1.

In the binary case, the *upper benchmark* is

$$\mathrm{UB}_N := \frac{\pi_0^{k(N_0)} (1-\pi_0)^{N_0 - k(N_0)} \pi_1^{k(N)-k(N_0)} (1-\pi_1)^{N_1 - (k(N)-k(N_0))}}{\pi^{k(N)} (1-\pi)^{N-k(N)}}, \tag{2}$$

where

$$\pi := \frac{N_0}{N} \pi_0 + \frac{N_1}{N} \pi_1 .$$

The upper benchmark is the final value $\mathrm{UB}_N = U_N$ of the likelihood ratio martingale

$$U_n := \begin{cases} \frac{\pi_0^{k(n)} (1-\pi_0)^{n-k(n)}}{\pi^{k(n)} (1-\pi)^{n-k(n)}} & \text{if } n \le N_0 \\ \frac{\pi_0^{k(N_0)} (1-\pi_0)^{N_0-k(N_0)} \pi_1^{k(n)-k(N_0)} (1-\pi_1)^{(n-N_0)-(k(n)-k(N_0))}}{\pi^{k(n)} (1-\pi)^{n-k(n)}} & \text{otherwise,} \end{cases} \tag{3}$$

where $n = 0, \ldots, N$. Unlike (1), (3) easily extends to other statistical models. Some of the standard statistical models are closed under convex closure, and for them the upper benchmark has a particularly simple expression.

The trajectory of the likelihood ratio martingale (3) is shown as the yellow line in Figure 2. It is close to a straight line, which makes it look very different from the lower benchmark. If, instead, we showed UB_n (as defined in (2) with n in place of N) versus $n > N_0$, the lines for the two benchmarks would be indistinguishable. Figure 2 only shows that the final values are close (in numbers, they are 7.6×10^{268} and 3.1×10^{269}). However, the line $n \mapsto \mathrm{UB}_n$ would be difficult to interpret.

The last two boxplots in Figure 3 show the median and the quartiles of the empirical distributions over 10^6 simulations for the two benchmarks, and their whiskers show the 5% and 95% quantiles. The boxplots are notched, with the notches indicating confidence intervals for the median (with this large number of simulations, the confidence intervals are

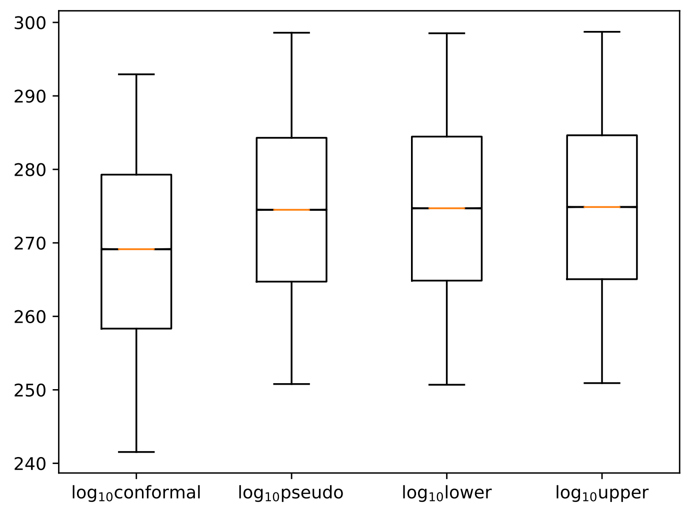

Figure 3: The boxplots over 10^6 simulations for the $\log_{10}$ of the final values of the custom-made conformal test martingale ("$\log_{10}$ conformal"), the corresponding conformal e-pseudomartingale ("$\log_{10}$ pseudo"), the lower benchmark ("$\log_{10}$ lower"), and the upper benchmark ("$\log_{10}$ upper"), as described in text.

very narrow; a less extreme case with visible notches will be shown in Figure 6). These two boxplots are very similar, and the medians in them are approximately $10^{274.71}$ and $10^{274.88}$.

The following proposition says that the final values of the upper and lower benchmarks are fairly close to each other asymptotically.

Proposition 3 *As $N_0 \to \infty$ and $N_1 \to \infty$,*

$$\frac{2N\pi(1-\pi)}{N_0\pi_0(1-\pi_0)+N_1\pi_1(1-\pi_1)} \ln \frac{\mathrm{UB}_N}{\mathrm{LB}_N} \xrightarrow{\text{law}} \xi^2, \tag{4}$$

where $\xi \sim N(0,1)$.

Informally, (4) implies

$$\log_{10} \frac{\mathrm{UB}_N}{\mathrm{LB}_N} \approx \frac{N_0\pi_0(1-\pi_0)+N_1\pi_1(1-\pi_1)}{2N\pi(1-\pi)\ln 10}\xi^2 \le \frac{\xi^2}{2\ln 10}, \tag{5}$$

where $\approx$ is used to signify the approximate equality of distributions, and the inequality follows from Jensen's inequality applied to the concave function $\pi \in [0,1] \mapsto \pi(1-\pi)$. Figure 4 shows the distributions of $\log_{10}(\mathrm{UB}_N / \mathrm{LB}_N)$, its approximation as given by the expression following $\approx$ in (5), and its upper bound as given by the expression following $\le$ in (5). We can see that the number of observations $N = 10^4$ (split in half by the changepoint) is sufficient for the asymptotic approximation to work. The median in the column "simulation" is approximately 0.085, and so the difference between the two benchmarks will not typically be noticeable on our plots.

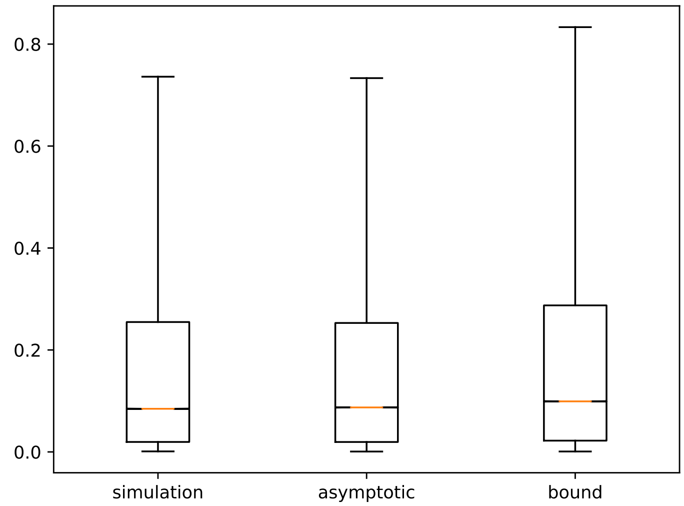

Figure 4: The decimal logarithm of $\mathrm{UB}_N \,/\, \mathrm{LB}_N$ in the model situation, its asymptotic approximation, and an upper bound for it, as described in text, based on 10^6 simulations.

4. Custom-made conformal test martingales

In this section we will discuss conformal test martingales specifically adapted to detecting changepoints. As in the previous section, and until Section 6, we use $B(0.1)$ as the pre-change distribution and $B(0.4)$ as the post-change distribution. The number of observations is 10^4 and the changepoint is in the middle of the dataset, so that the first 5000 observations are generated from $B(0.1)$ and the remaining 5000 from $B(0.4)$.

For a detailed definition of conformal test martingales, see, e.g., Vovk (2021) and Vovk et al. (2021). What follows is a brief reminder focusing on the main ideas. As usual, we start from a nonconformity measure A. In the case of conformal testing, a successful A does not have to be a good measure of how badly, or how well, a new observation conforms to a given multiset of observations; e.g., the Simple Jumper martingale (Vovk et al., 2021) used in Section 2 does not change if we use $-A$ in place of A. An input stream of observations z_n is transformed into a stream of (smoothed) p-values p_n as usual:

$$p_n := \frac{|\{i : \alpha_i > \alpha_n\}| + \theta_n \,|\{i : \alpha_i = \alpha_n\}|}{n}, \tag{6}$$

where i ranges over $\{1, \dots, n\}$, $\alpha_1, \dots, \alpha_n$ are the nonconformity scores for $z_1, \dots, z_n$ computed using A, and θ_n are random numbers distributed uniformly on the interval $[0, 1]$ (all independent).

The standard property of validity for conformal prediction (Vovk et al., 2005, Proposition 2.8) is that the p-values (6) are independent and distributed uniformly on $[0, 1]$. This way we turn our composite null hypothesis (the IID assumption) into a simple null hypothesis (uniformity) about the p-values. The next step is to gamble against the uniformity of

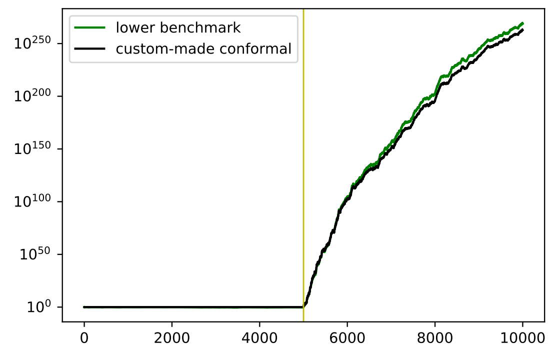

Figure 5: The custom-made conformal test martingale and the lower benchmark, as described in text.

the p-values using *betting functions*, i.e., functions $f : [0, 1] \to [0, \infty]$ that integrate to 1. In conformal testing, at step n a betting function f_n is chosen (in a measurable manner) with the knowledge of the first $n-1$ p-values $p_1, \dots, p_{n-1}$. The product $S_n := f_1(p_1) \dots f_n(p_n)$, $n = 0, 1, \dots$ (with $S_0 := 1$), is the corresponding *conformal test martingale*. It is interpreted as the capital of a gambler playing against the null hypothesis, and S_n represents the amount of evidence found against the null hypothesis by time n. Our game is fair (under the null hypothesis) in that the expected value of S_n given the history $p_1, \dots, p_{n-1}$ up to time $n-1$ equals the capital S_{n-1} at that time.

Conformal test martingales are *exchangeability martingales*, i.e., satisfy

$$\mathbb{E}(S_n \mid S_1, \dots, S_{n-1}) = S_{n-1} \tag{7}$$

under any exchangeable distribution on the observations. By de Finetti's theorem, in this context the assumption of exchangeability is equivalent to the IID assumption under the weak condition that the observation space is Borel (which is satisfied in applications).

Next let us find a conformal test martingale that is expected to work well under the true data distribution. Our argument will be somewhat informal. During the first N_0 trials we do not gamble, so let us consider a trial $n > N_0$. Taking the identity function as the nonconformity measure (the difference between conformity and nonconformity is essential in this context), by (6) we obtain a p-value $p_n \in [0, k(n)/n]$ with probability π_1, and we obtain $p_n \in [k(n)/n, 1]$ with probability $1 - \pi_1$. Since the expected value of $k(n)/n$ is $(N_0\pi_0 + (n - N_0)\pi_1)/n$, the likelihood ratio betting function

$$f_n(p) := \begin{cases} \frac{n\pi_1}{N_0\pi_0+(n-N_0)\pi_1} & \text{if } p \le \frac{N_0\pi_0+(n-N_0)\pi_1}{n} \\ \frac{n(1-\pi_1)}{N_0(1-\pi_0)+(n-N_0)(1-\pi_1)} & \text{otherwise} \end{cases} \tag{8}$$

is in some sense optimal, as shown in Fedorova et al. (2012, Theorem 2). The black line in Figure 5 shows the trajectory of the corresponding conformal test martingale.

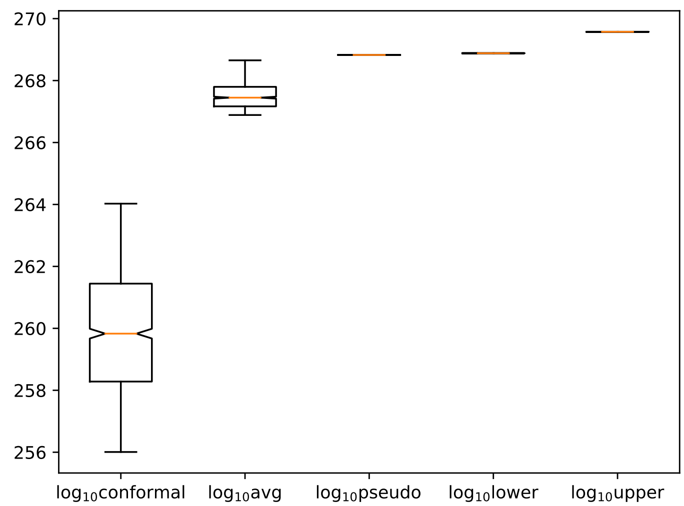

Figure 6: The analogue of Figure 3 for a fixed dataset (corresponding to the seed 2021 of the NumPy pseudorandom number generator) with an extra boxplot $\log_{10}$ avg (average over 10^6 runs) explained in text. The number of simulations is decreased to 10^3.

The betting functions (8) involve the expected value of $k(n)/n$. We can often improve the performance of the conformal test martingale shown in Figure 5 if we replace (8) by

$$f_n(p) := \begin{cases} \frac{n\pi_1}{k(n)} & \text{if } p \le \frac{k(n)}{n} \\ \frac{n(1-\pi_1)}{n-k(n)} & \text{otherwise.} \end{cases} \tag{9}$$

However, the resulting process is not a genuine martingale but a conformal e-pseudomartingale, in the terminology of Vovk (2020a).

In plots such as Figure 5 the trajectories of the two benchmarks, conformal e-pseudomartingale, and the custom-made conformal martingale look very close, but in fact the difference between the final values of those processes can often be as large as 10^{10}-fold. The boxplot "$\log_{10}$ conformal" in Figure 3 corresponds to the black line in Figure 5 (which represents the first simulation out of the 10^6 represented in the boxplot), the boxplot "$\log_{10}$ lower" corresponds to the green lines in Figures 2 and 5, and the boxplot "$\log_{10}$ upper" corresponds to the yellow lone in Figure 2. The boxplot "$\log_{10}$ pseudo" gives statistics for the final values of the conformal e-pseudomartingale based on (9), whose plot is not shown but would have been indistinguishable from the green line in Figure 5. In numbers, the medians for the final values of the four processes in the order in which they are shown in Figure 3 (which is the ascending order) are, approximately, $10^{269.14}$, $10^{274.50}$, $10^{274.71}$, and $10^{274.88}$ (the last two numbers were already given above).

The boxplot for the conformal test martingale in Figure 3 is slightly longer than the other three boxplots. The explanation is that conformal test martingales are randomized (because of the dependence of (6) on θ_n), unlike, e.g., the lower benchmark process. The corresponding boxplots for a fixed dataset (the same one that was used in Figures 1, 2, and 5) are shown in Figure 6, along with an extra boxplot labelled $\log_{10}$ avg, to be explained momentarily.

It appears from Figure 6 that, for a fixed dataset, the final values of the conformal e-pseudomartingales are constant a.s. This is indeed the case: e.g., with probability one under any IID measure, the condition $p \le k(n)/n$ in (9) holds for $p = p_n$ if and only if the nth observation is 1.

On the other hand, the final value of the conformal test martingale in Figure 6 is very volatile, with the upper quartile around 10^3 times larger than the lower quartile. An easy way to decrease the volatility of a randomized test martingale is to average its trajectory over a number of independent runs (as explained in Vovk 2020b in the context of e-variables); normally, the result will still be a valid test martingale. The results of averaging the conformal test martingale over 10^6 runs are shown in the new boxplot labelled $\log_{10}$ avg. The operation of averaging not only reduces volatility but also greatly improves the typical performance, the reason being that on the log scale the average of vastly different numbers is close to their maximum. The first two boxplots in Figure 6 are based on 10^3 simulations of the conformal test martingale (for the first boxplot) or averaged conformal test martingale (for the second one).

Unfortunately, in the case of averaging conformal test martingales there is no guarantee that the average will still be an exchangeability martingale, since different conformal test martingales involve different filtrations (Vovk, 2021, Remark 3.3). And indeed, in Section 7 we will see an example where the average is not an exchangeability martingale.

5. More natural conformal test martingales

The martingales whose trajectories are shown in Figures 2–5 depend very much on the knowledge of the true data-generating mechanism. Can we obtain comparable results without blatant optimization (requiring such knowledge)? This is the topic of this section.

Let us generalize the betting function (8) to

$$f_{(a,b)}(p) := \begin{cases} \frac{b}{a} & \text{if } p \le a \\ \frac{1-b}{1-a} & \text{otherwise,} \end{cases} \tag{10}$$

where $a, b \in (0,1)$. It is easy to see that $\int f_{(a,b)} = 1$. Apart from the betting functions (10) we will use the trivial function $f_\square$, $f_\square(p) := 1$ for all p. Let S_n be the conformal test martingale

$$S_n := \int f_{s_1}(p_1) \dots f_{s_n}(p_n) \mu(\mathrm{d}(s_1, s_2, \dots)), \tag{11}$$

where $p_1, p_2, \dots$ is the underlying sequence of conformal p-values and μ is the distribution of the following Markov chain with states $s_1, s_2, \dots$.

The Markov chain is defined in the spirit of tracking the best expert in prediction with expert advice (Herbster and Warmuth, 1998; Vovk, 1999). The state space is $\{\square\} \cup (0,1)^2$, and $R \in (0,1)$ is the parameter (typically a small number). The initial state is $s_1 := \square$ (the *sleeping* state). The transition function is:

Algorithm 1: Sleeper/Stayer

Data: p-values $p_1, p_2, \dots$
Result: conformal test martingale $S_0, S_1, S_2, \dots$
$S_0 := S_\Box := 1$;
for $(a,b) \in \mathbf{G}^2$ **do**
  $S_{a,b} := 0$
end
for $n = 1, 2, \dots$ **do**
  for $(a,b) \in \mathbf{G}^2$ **do**
    $S_{a,b} := S_{a,b} f_{(a,b)}(p_n)$
  end
  $S_n := S_\Box + \sum_{(a,b)\in\mathbf{G}^2} S_{a,b}$;
  for $(a,b) \in \mathbf{G}^2$ **do**
    $S_{a,b} := S_{a,b} + R S_\Box/(G-1)^2$
  end
  $S_\Box := (1-R) S_\Box$
end

- if the current state is $\Box$, with probability $1-R$ the state remains $\Box$, and with probability R a new state (a,b) is chosen from the uniform distribution in $(0,1)^2$;
- the states $(a,b) \in (0,1)^2$ are absorbing: if the current state is $(a,b) \in (0,1)^2$, it will stay (a,b).

In our implementation of the procedure (11), we replace the square $(0,1)^2$ by the grid $\mathbf{G}^2$, where

$$\mathbf{G} := \left\{ \frac{1}{G}, \frac{2}{G}, \dots, \frac{G-1}{G} \right\} \tag{12}$$

and G (positive integer) is another parameter. The resulting procedure is shown as Algorithm 1.

The intuition behind Algorithm 1 is that, in order to gamble against the uniformity of $(p_1, p_2, \dots)$, we distribute our initial capital of 1 among accounts $S_{a,b}$ indexed by $(a,b) \in \mathbf{G}^2$, and there is also a sleeping account $S_\Box$. We start from all money invested in the sleeping account, but at the end of each step a fraction R of that money is moved to the active accounts $S_{a,b}$ and divided between them equally. On account $S_{a,b}$ we gamble against the uniformity of the input p-values using the betting function $f_{(a,b)}$.

Figure 7 (the line in cyan) suggests that we can improve on the result of Figure 1 using a fairly natural, and in fact very basic, conformal test martingale. In Figure 7 we use the identity nonconformity measure and the Sleeper/Stayer betting martingale of Algorithm 1, and the parameters are $R := 0.001$ and $G := 10$; therefore, a and b are chosen from the grid $\{0.1, 0.2, \dots, 0.9\}$. The final value of the resulting conformal test martingale is closer (on the log scale) to those in Figure 5 than in Figure 1.

To improve further the performance of a natural conformal test martingale, let us make another step towards the custom-made martingale (8). The new martingale will be defined

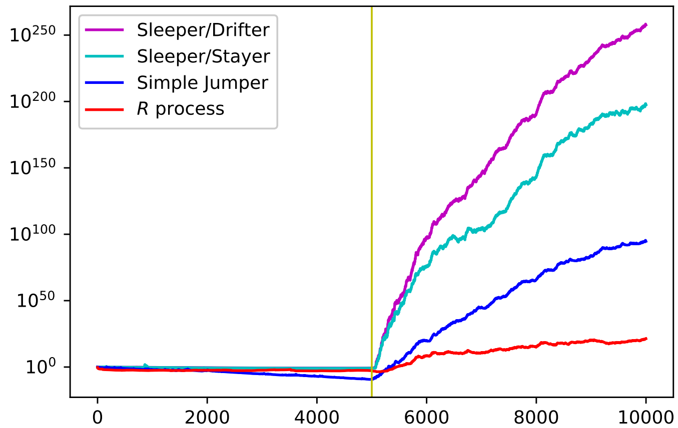

Figure 7: Various conformal test martingales and the R process (Ramdas et al., 2021), as described in text; the final values are approximately 2.3×10^{21} (R process), 4.7×10^{94} (Simple Jumper), 2.8×10^{197} (Sleeper/Stayer), and 4.6×10^{257} (Sleeper/Drifter).

as an average of the following "expert martingales". An expert martingale is characterized by a vector parameter $(N_0, \pi_0, \pi_1) \in \{1, 2, \dots\} \times (0, 1)^2$ and is the custom-made martingale (8) for these postulated (N_0, π_0, π_1), rather than the unknown real ones. (In this and the next paragraphs, we will use N_0, π_0, π_1 as local variables; in the end they will be integrated out, and we will again be able to use them in the global sense introduced in Section 2.) The expert sleeps (does not gamble) until time N_0, and at each time $n > N_0$ it uses the betting function (8). This betting function is of the form (10) with $b := \pi_1$ and $a = a_n$ being the weighted average of π_0 and π_1 with the weights N_0/n and $1 - N_0/n$, respectively. Therefore, a_n gradually drifts from π_0 towards π_1.

The *Sleeper/Drifter martingale* depends on three parameters: G, determining the grid (12), M ($M := 1$ is a good value, but larger values of M improve computational efficiency), and R (the rate at which the experts, who are originally sleeping, wake up). It is the average of the experts w.r. to the following probability measure:

- all three parameters are independent;
- $N_0 = iM$, where $i \in \{1, 2, \dots\}$ is generated according to the geometric distribution with parameter RM;
- π_0 and π_1 are generated from the uniform distribution in the grid (12).

The overall procedure is given as Algorithm 2. The key array in this algorithm is $(S_{i,a,b})$, where $S_{i,a,b}$ is the total capital of the experts drifting from a towards b who woke up at time iM. Now we can say that $S_\square$ is the total capital of the experts who are still asleep; as an expert wakes up, its capital moves from $S_\square$ to one of the $S_{i,a,b}$.

The performance of Algorithm 2 is shown as the magenta line in Figure 7. The parameters used there are $G = 10$, $M = 100$, and $R = 0.001$. (There is not much sensitivity to

Algorithm 2: Sleeper/Drifter

Data: p-values $p_1, p_2, \ldots$
Result: conformal test martingale $S_0, S_1, S_2, \ldots$
$S_0 := S_\square := 1$;
for $i = 1, 2, \ldots$ *and* $(a, b) \in \mathbf{G}^2$ **do**
  $S_{i,a,b} := 0$
end
for $n = 1, 2, \ldots$ **do**
  for $i < n/M$ *and* $(a, b) \in \mathbf{G}^2$ **do**
    $a' := \frac{iM}{n} a + \left(1 - \frac{iM}{n}\right) b$;
    $S_{i,a,b} := S_{i,a,b} f_{(a',b)}(p_n)$
  end
  $S_n := S_\square + \sum_{(i,a,b)\in\{1,2,\ldots\}\times\mathbf{G}^2} S_{i,a,b}$;
  if *n is divisible by M* **then**
    for $(a, b) \in \mathbf{G}^2$ **do**
      $S_{n/M,a,b} := RMS_\square/(G-1)^2$
    end
    $S_\square := (1 - RM)S_\square$
  end
end

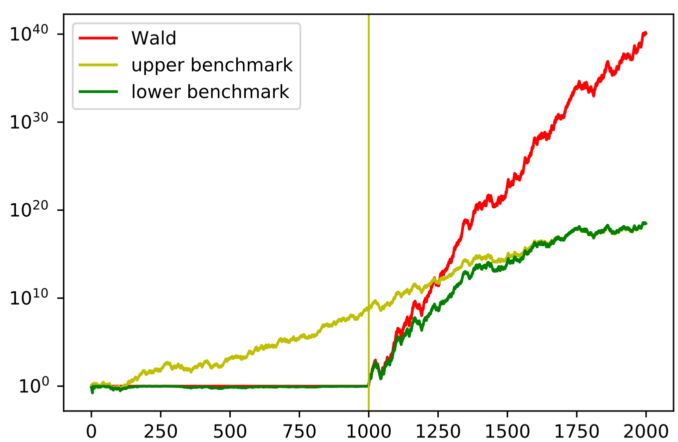

Figure 8: The analogue of Figure 2 (left panel) for the medium scenario.

the values of the parameters; e.g., if we decrease R to 10^{-4} or 10^{-5}, we will get final values of about the same order of magnitude: 4.9×10^{258} or 7.6×10^{257}, respectively.)

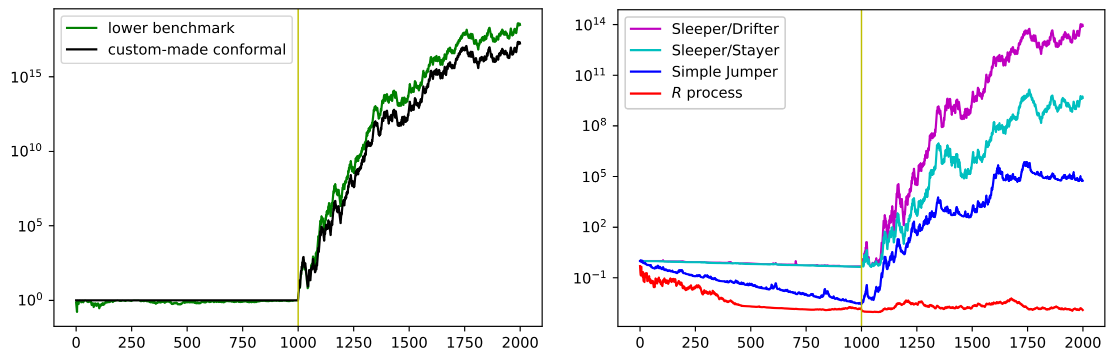

Figure 9: The analogue of Figures 5 (shown as the left panel) and 7 (shown as the right panel) in the medium scenario.

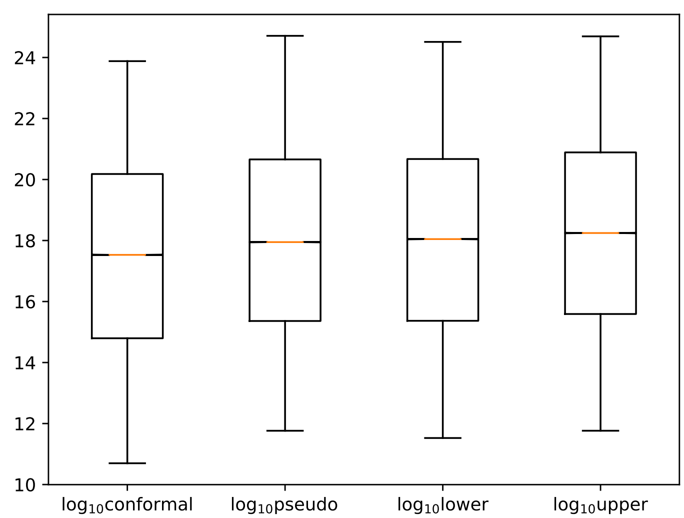

Figure 10: The analogue of Figure 3 for the medium scenario, with the number of simulations still 10^6.

6. Smaller datasets

In this section we will consider two less extreme scenarios, which we will label as medium and small (and will refer to the scenario of the previous sections as *large*). In the *medium* scenario, 1000 observations from $B(0.3)$ are followed by 1000 observations from $B(0.5)$. Figures 8–10 are analogues for the medium scenario of some figures in the previous sections and exhibit similarities with the large scenario.

In the *small* scenario, 100 observations from $B(0.2)$ are followed by 1000 observations from $B(0.5)$. The dependence on the choice of parameters for conformal test martingales

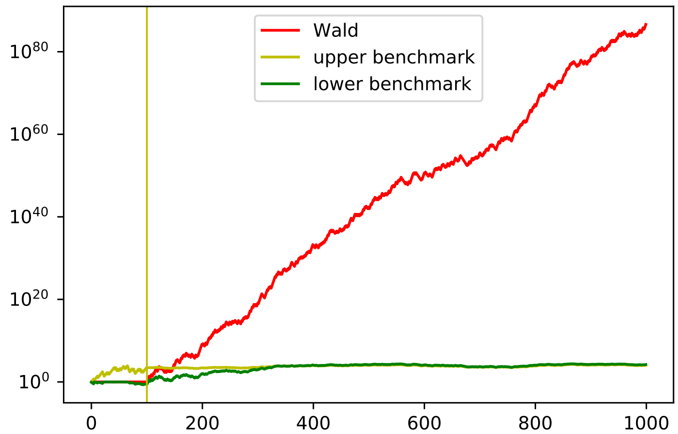

Figure 11: The analogue of Figure 2 for the small scenario.

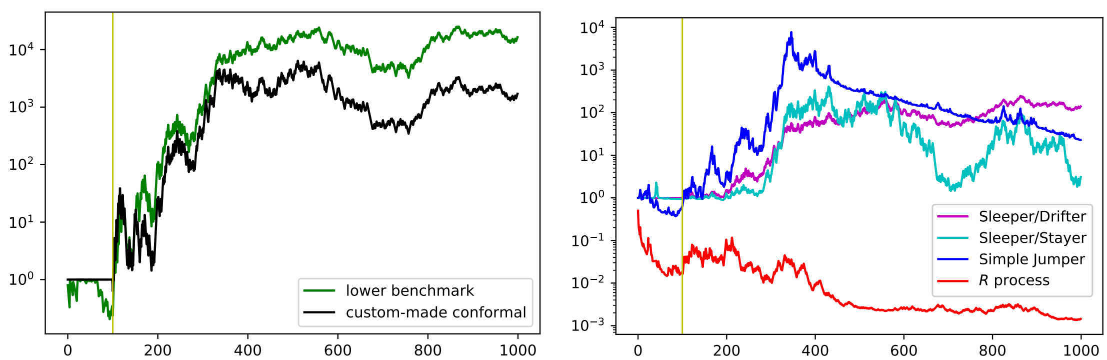

Figure 12: The analogue of Figures 5 (shown as the left panel) and 7 (shown as the right panel) in the small scenario.

becomes much more pronounced, but we keep all old values for the parameters of the Sleeper/Stayer and Sleeper/Drifter (even though other values may improve their performance significantly). One difference from the results for the large and medium scenarios is the improved performance of the Simple Jumper as compared with the Sleeper/Stayer and Sleeper/Drifter. Another difference is that, since most of the observations in the small scenario are post-change, we can clearly see that all martingales, and especially the Simple Jumper, at some point start losing evidence. Possible ways of preventing heavy loss of evidence are discussed in Shafer and Vovk (2019, Chapter 11).

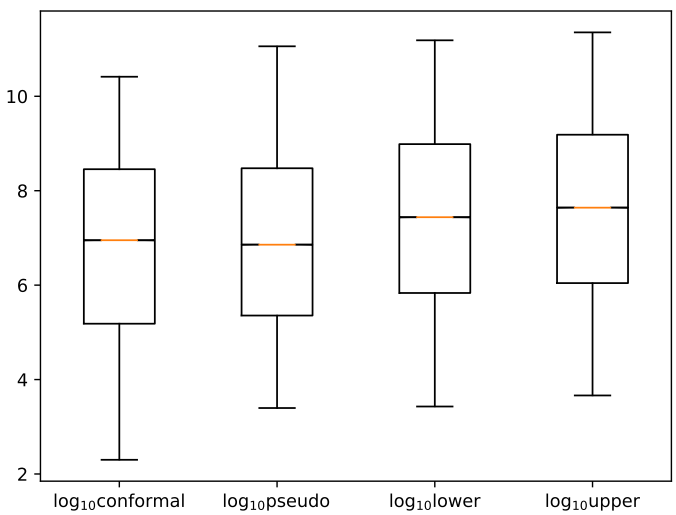

Figure 13: The analogue of Figure 3 for the small scenario, with the number of simulations still 10^6.

7. Testing the validity of putative test martingales

The performance of some of the conformal test martingales constructed in this paper might appear too good, and some of our processes are not guaranteed to be exchangeability martingales (such as the average process of Figure 6). Therefore, it may be useful to be able to test whether such processes are martingales in simulation studies (of course, we have theoretical guarantees of validity for conformal test martingales, but even for them mistakes in implementation are always possible). The testing method of this section will use the following large deviations inequality based on Doléans's supermartingale of Shafer and Vovk (2019, Section 3.2), which we first give in terms of e-values (Vovk and Wang, 2021) and p-values. The defining property of an e-value is that it is nonnegative and its expected value is at most one; a large e-value is interpreted as evidence against our postulated stochastic mechanism (the null hypothesis).

Proposition 4 *Let $F_1, \dots, F_K$, $K \ge 4$, be independent nonnegative random variables with expected value 1, and let M be a positive integer. Then*

$$e := \frac{1}{M} \sum_{m=1}^{M} \exp\left(K^{1-m/2M} (\bar{F} - 1) - K^{-m/M} \sum_{k=1}^{K} (F_k - 1)^2 \right), \tag{13}$$

where $\bar{F} := \frac{1}{K} \sum_{k=1}^{K} F_k$ is the average of the F_k, is a valid e-value, and $\frac{1}{e} \wedge 1$ is a valid p-value.

Proof The statement about (13) being an e-value follows from the right-hand side of (13) being the final value of a test supermartingale (i.e., a nonnegative supermartingale with

(π_0, π_1)	(N_0, N_1)	K	mean	bound	median	quartiles
$(0.1, 0.4)$	$(10, 10)$	10^9	0.99993	1.00054	0.33016	$[0.13964, 0.84562]$
$(0.4, 0.5)$	$(10, 10)$	10^9	1.00000	1.00008	0.89615	$[0.66667, 1.21212]$
$(0.4, 0.5)$	$(100, 100)$	10^9	0.99985	1.00040	0.36630	$[0.14232, 0.94952]$

Table 1: The mean $\frac{1}{K}\sum_k F_k$, its upper bound in (15), and the median and interquartile range of $F_1, \dots, F_K$.

initial value 1), namely an average of Doléans supermartingales (Shafer and Vovk, 2019, Proposition 3.4). The statement about $\frac{1}{e} \wedge 1$ being a p-value follows from $e \mapsto \frac{1}{e} \wedge 1$ being an e-to-p calibrator (Vovk and Wang, 2021, Proposition 2.2). ■

In the main part of this section we will use Proposition 4 in the form of the following inequality.

Corollary 5 *Let $F_1, \dots, F_K$, $K \geq 4$, be independent nonnegative random variables with expected value 1, let M be a positive integer, and let $\epsilon > 0$. Define $X > 0$ as the only solution to*

$$\sum_{m=1}^{M} \exp\left(K^{1-m/2M} X - K^{-m/M} \sum_{k=1}^{K} (F_k - 1)^2\right) = \frac{M}{\epsilon} \tag{14}$$

(the left-hand side is strictly increasing in X). Then

$$\mathbb{P}\left(\frac{1}{K}\sum_{k=1}^{K} F_k < 1 + X\right) \geq 1 - \epsilon. \tag{15}$$

Proof If the inner inequality in (15) is violated, we will have

$$\frac{1}{M}\sum_{m=1}^{M} \exp\left(K^{-m/2M} \sum_{k=1}^{K} (F_k - 1) - K^{-m/M} \sum_{k=1}^{K} (F_k - 1)^2\right) \geq \frac{1}{\epsilon}$$

instead of (14). The probability of this event is at most ϵ since the reciprocal to (13) is a p-value. ■

Let us use $M := 5$. For a few sets of values for (N_0, N_1) and (π_0, π_1), Table 1 gives some statistics for the final values F_k of the custom-made conformal test martingale with the betting functions (8) designed for the pre-/post-change parameters (π_0, π_1) but run on the IID data with parameter π_0; the numbers of pre- and post-change observations is N_0 and N_1 respectively. The closeness of the means and bounds to 1 suggests that the processes are really test martingales. Of course, the bound is never exceeded by the actual mean.

Table 2 is analogous to Table 1 but gives statistics for the average over 10^3 runs of conformal test martingales. The means are still close to 1 and do not exceed the bounds. Unfortunately, this kind of statistics does not allow us to check deviations of the average

(π_0, π_1)	(N_0, N_1)	K	mean	bound	median	quartiles
$(0.1, 0.4)$	$(10, 10)$	10^6	0.99894	1.00570	0.67879	$[0.38007, 1.37617]$
$(0.4, 0.5)$	$(10, 10)$	10^6	1.00007	1.00207	0.94866	$[0.74567, 1.15930]$
$(0.4, 0.5)$	$(100, 100)$	10^6	0.99972	1.00994	0.43602	$[0.17872, 1.06452]$

Table 2: The analogue of Table 1 for the average of the conformal test martingale over 10^3 runs.

K^*	K	A	mean	bound	median	quartiles
10^6	482,311	10^3	1.00426	1.00101	0.99580	$[0.89924, 1.00682]$
10^9	400,000,071	1	1.00001	1.00007	0.83333	$[0.83333, 1.42857]$
10^9	447,299,138	10	1.00266	1.00005	0.96172	$[0.88585, 1.06718]$
10^9	470,992,540	10^2	1.00353	1.00005	0.98566	$[0.91111, 1.02118]$
10^9	482,226,950	10^3	1.00452	1.00004	0.99589	$[0.89931, 1.00684]$

Table 3: Statistics for the conditional validity of the average conformal test martingale with $(\pi_0, \pi_1) = (0.1, 0.4)$, as described in text.

conformal test martingale from being a martingale, since the expectation of the final value of the average is still 1.

The method that we have used so far can be easily adapted for the purpose of checking the martingale property, and it will show that the average conformal test martingale is not a martingale itself (under the null hypothesis). Let S_n be an average conformal test martingale; it will be assumed positive. The defining property of a martingale is (7). The method that we have used tests the crude implication $\mathbb{E}(S_n) = 1$ of the defining property, which we know to hold for an average of martingales; the modification will test $\mathbb{E}(S_n \mid S_{n-1}) = S_{n-1}$, i.e., $\mathbb{E}(S_n/S_{n-1} \mid S_{n-1}) = 1$.

Table 3 summarizes a case where $\mathbb{E}(S_n/S_{n-1} \mid S_{n-1} \geq 1) > 1$ (so that S possesses a momentum: a rise in the value of S creates a tendency to a further rise). The conformal test martingale is the one with the betting functions (8), where $N_0 := 2$ and $(\pi_0, \pi_1) = (0.1, 0.4)$; it is averaged over A simulations. The value of K is the number of runs of the average conformal test martingale with $S_{n-1} \geq 1$, where $n := 5$. These runs are selected from K^* runs by discarding the runs leading to $S_{n-1} < 1$. The mean, median, and quartiles are those of S_n/S_{n-1} over the K selected runs, and the bound is as given by Corollary 5 with $\epsilon := 0.01$. We can see that the bound is exceeded by the actual mean except for the case where $A = 1$ (and so there is no averaging). The mean mostly depends on A, and the bound on K.

To get an idea of how serious the violation of the bounds in Table 3 is, we can apply Proposition 4 directly. The p-values computed using Proposition 4 from Table 3 are tiny, except, of course, for the second row, where the e-value is 0.25 and the p-value is 1. Even for the top row, the p-value is below 10^{-44}.

8. Further discussion

In this paper we have discussed only the case of binary observations, in which the simple betting functions (10) are appropriate. This can be regarded as a first step of an interesting research programme. We can simulate different model situations that can be analyzed theoretically and develop suitable conformal test martingales, as we did in this paper for a binary model situation. Perhaps the next in line are the Gaussian model with a constant variance and a change in the mean, the Gaussian model with a constant mean and a change in the variance, and the exponential model (as in, e.g., Wald 1947, Part II, and Tartakovsky et al. 2015). Custom-made conformal test martingales (such as those in Section 4) provide clear goals for more natural conformal test martingales, and even give ideas of how these goals can be attained. These ideas, in turn, add to the toolbox that we can use for dealing with practical problems, where we often have only a vague notion of the true data-generating distribution. See Nouretdinov et al. (2021) for some results in this direction.

Even in the case of binary observations, better conformal martingales can be designed. The function (8) is discontinuous, and it leads to a drop in its performance: when p is close to the borderline value $\frac{N_0\pi_0+(n-N_0)\pi_1}{n}$, it is better not to gamble at all than to use (8).

Acknowledgments

Thanks to Aaditya Ramdas for helpful suggestions, to Ivan Petej, Ilia Nouretdinov, and Alex Gammerman for illuminating discussions, and to the three anonymous referees of the conference version of this paper for useful comments including suggestions for further research. This work has been supported by Amazon and Stena Line.

References

Valentina Fedorova, Ilia Nouretdinov, Alex Gammerman, and Vladimir Vovk. Plug-in martingales for testing exchangeability on-line. In John Langford and Joelle Pineau, editors, *Proceedings of the Twenty Ninth International Conference on Machine Learning*, pages 1639–1646. Omnipress, 2012.

R. K. Getoor and M. J. Sharpe. Conformal martingales. *Inventiones Mathematicae*, 16: 271–308, 1972.

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.

Mark Herbster and Manfred K. Warmuth. Tracking the best expert. *Machine Learning*, 32:151–178, 1998.

Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, New York, third edition, 2005.

Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A*, 231:289–337, 1933.

Ilia Nouretdinov, Vladimir Vovk, and Alex Gammerman. Conformal changepoint detection in continuous model situations. *Proceedings of Machine Learning Research*, 152, 2021. COPA 2021, to appear.

Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen. How can one test if a binary sequence is exchangeable? Fork-convex hulls, supermartingales, and Snell envelopes. Technical Report arXiv:2102.00630 [math.ST], arXiv.org e-Print archive, February 2021.

Glenn Shafer. The language of betting as a strategy for statistical and scientific communication (with discussion). *Journal of the Royal Statistical Society A*, 184:407–478, 2021.

Glenn Shafer and Vladimir Vovk. *Game-Theoretic Foundations for Probability and Finance*. Wiley, Hoboken, NJ, 2019.

Alexander Tartakovsky, Igor Nikiforov, and Michèle Basseville. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. CRC Press, Boca Raton, FL, 2015.

Vladimir Vovk. Derandomizing stochastic prediction strategies. *Machine Learning*, 35: 247–282, 1999.

Vladimir Vovk. Conformal e-prediction for change detection. Technical Report arXiv:2006.02329 [math.ST], arXiv.org e-Print archive, June 2020a.

Vladimir Vovk. A note on data splitting with e-values: online appendix to my comment on Glenn Shafer's "Testing by betting". Technical Report arXiv:2008.11474 [stat.ME], arXiv.org e-Print archive, August 2020b. This is part of a comment on Shafer (2021).

Vladimir Vovk. Testing for concept shift online. Technical Report arXiv:2012.14246 [cs.LG], arXiv.org e-Print archive, December 2020c.

Vladimir Vovk. Testing randomness online. *Statistical Science*, 2021. To appear, published online.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination, and applications. *Annals of Statistics*, 49:1736–1754, 2021.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

Vladimir Vovk, Ivan Petej, Ilia Nouretdinov, Ernst Ahlberg, Lars Carlsson, and Alex Gammerman. Retrain or not retrain: Conformal test martingales for change-point detection. *Proceedings of Machine Learning Research*, 152, 2021. COPA 2021, to appear.

Abraham Wald. *Sequential Analysis*. Wiley, New York, 1947.

Abraham Wald and Jacob Wolfowitz. Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics*, 19:326–339, 1948.

John B. Walsh. A property of conformal martingales. *Séminaire de probabilités (Strasbourg)*, 11:490–492, 1977.