Proceedings of Machine Learning Research 179:1–12, 2022 Conformal and Probabilistic Prediction and Applications

Conformal testing: binary case with Markov alternatives

Vladimir Vovk V.VOVK@RHUL.AC.UK
Ilia Nouretdinov ILIA.NOURETDINOV@RHUL.AC.UK
Alex Gammerman A.GAMMERMAN@RHUL.AC.UK
Royal Holloway, University of London, Egham, Surrey, UK

Editor: Ulf Johansson, Henrik Boström, Khuong An Nguyen, Zhiyuan Luo, and Lars Carlsson

Abstract

We continue study of conformal testing in binary model situations. In this paper we consider Markov alternatives to the null hypothesis of exchangeability. We propose two new classes of conformal test martingales; one class is statistically efficient in our experiments, and the other class partially sacrifices statistical efficiency to gain simplicity and computational efficiency.

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

1. Introduction

This paper treats a problem similar to the one considered in Ramdas et al. (2022): we would like to test online the null hypothesis of exchangeability of binary observations under Markov alternatives. By de Finetti's theorem, the null hypothesis is equivalent to the observations being independent and identically distributed (IID).

The simplest way of online hypothesis testing is to use *test martingales*, which are defined as nonnegative processes with initial value 1 that are martingales under the null hypothesis; see, e.g., Shafer et al. (2011). Such processes, for the null hypothesis of exchangeability, can be constructed using the method of conformal prediction (Vovk et al., 2022), and we will refer to them as conformal test martingales. A previous paper (Vovk, 2021b) constructs custom-made conformal test martingales for different alternative hypotheses, those of a changepoint.

The method of Ramdas et al. (2022), which is specifically devoted to Markov alternatives, is more general: instead of a test martingale the authors construct a "safe e-process" (to be defined in the next section). Safe e-processes are closely related to test martingales and admit a similar interpretation as the capital of a gambler trying to discredit the null hypothesis. Our methods give similar results to the methods of Ramdas et al. (2022) in the model situations that we consider (following Ramdas et al. 2022). The advantage of our methods is that they extend easily to the usual setting of machine learning, where the observations are pairs (x, y) consisting of a potentially complex object x and its label y. In this usual setting the methods of Ramdas et al. (2022) do not work at all: see, e.g., Vovk (2021b, Remark 2); we will also discuss it briefly in Section 6.

In this paper we only design conformal test martingales for a simple Markov alternative hypothesis (a specific Markov probability measure). This is different from Ramdas et al. (2022), who are interested in testing against the composite alternative Markov hypothesis.

As in Ramdas et al. (2022), we could mix our conformal test martingales over all possible alternative hypotheses analytically (this is done in Vovk et al. 2022, Chapter 8). A simpler solution is to consider a dense grid of parameter values and use the arithmetic average of our conformal test martingales over the grid.

Remark 1 There is an unlikely possibility of confusion between the conformal test martingales used in this paper with the older notion of conformal martingales introduced in 1972 (Getoor and Sharpe 1972; see also Revuz and Yor 1999, Section 5.2). In this paper we will never use the expression "conformal martingale" (without "test") outside this remark.

2. Model situations

This section introduces the model situations considered in this paper, following Ramdas et al. (2022, Section 4.2). Our data consist of binary (0 or 1) observations generated from a known Markov model. We will use the notation $\text{Markov}(\pi_{1|0}, \pi_{1|1})$ for the probability distribution of a Markov chain with the transition probabilities $\pi_{1|0}$ for transitions $0 \to 1$ and $\pi_{1|1}$ for transitions $1 \to 1$; the probability that the first observation is 1 will always be assumed 0.5. The probability of the first observation plays a very minor role, and our null hypothesis is, essentially, that $\pi_{1|0} = \pi_{1|1}$.

In our computational experiments, we consider two cases and three scenarios. In the *hard case*, the model is $\text{Markov}(0.4, 0.6)$, and in the *easy case*, the model is $\text{Markov}(0.1, 0.9)$. (The hard case is harder to distinguish from the null hypothesis than the easy case.) The number of observations is $N := 10^4$ (as in Ramdas et al. 2022) or $N := 10^3$ or $N := 10^2$; we will refer to these scenarios as *large*, *medium*, and *small*, respectively.

In all our experiments we use 2022 as the seed for the NumPy pseudorandom number generator. (This, however, does not make the trajectories in our plots comparable between different scenarios.) The dependence on the seed will be explored in boxplots reported in

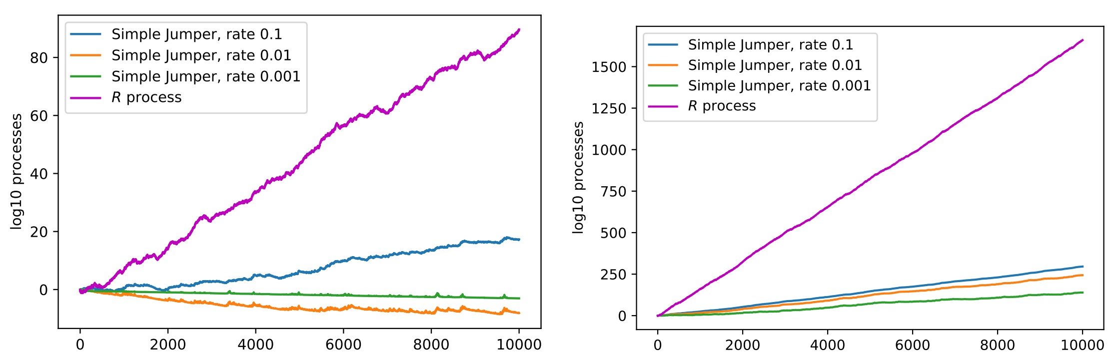

Figure 1: The process R of Ramdas et al. (2022) and the Simple Jumper in the large scenario. The vertical axis is logarithmic (base 10); e.g., for the process R we show the values $\log_{10} R$. Left panel: the hard case (the final value of R is about 10^{90}). Right panel: the easy case (the final value of R is about 10^{1660}).

Section 5; the seed affects not only the data but also the values of conformal test martingales, which are randomized processes, given the data.

Let B_π be the Bernoulli distribution on $\{0,1\}$ with parameter $\pi \in [0,1]$: $B_\pi(\{1\}) = \pi$. Set $\mathrm{Ber}(\pi) := B_\pi^\infty$. Our null hypothesis (the *IID model*, or, in our binary context, the *Bernoulli model*) is that the observations are generated from $\mathrm{Ber}(\pi)$ with unknown parameter π.

To detect deviations from the null hypothesis in the online mode (with the observations arriving sequentially), we typically use test martingales w.r. to $\mathrm{Ber}(\pi)$. See the previous paper (Vovk, 2021b) for the definitions.

Ramdas et al. construct a "safe e-process" $R = R_n$, which satisfies the following property: under any $\mathrm{Ber}(\pi)$, R is dominated by a test martingale $M_n^{(\pi)}$ w.r. to $\mathrm{Ber}(\pi)$, in the sense that $R_n \le M_n^{(\pi)}$ for all n and π (this property implies the definition of a safe e-process given by Ramdas et al.). The trajectories of the R process for the two cases, hard and easy, and for the large scenario are shown in Figure 1 (they coincide with those in Figure 4 in Ramdas et al. 2022 apart from using base 10 logarithms and a different randomly generated dataset). The figure also shows trajectories of the Simple Jumper martingale used earlier (see, e.g., Vovk 2021a) for various values of its parameter, called the jumping rate; it performs poorly in this context.

3. Two benchmarks

In this section we will discuss possible benchmarks that we can use for evaluating the quality of our conformal test martingales. Our goal will be to perform almost as well as the benchmarks. The *upper benchmark* is

$$\mathrm{UB}_n := \frac{\mathrm{Markov}(\pi_{1|0}, \pi_{1|1})([z_1, \dots, z_n])}{\mathrm{Ber}(0.5)([z_1, \dots, z_n])}, \tag{1}$$

where $[z_1, \dots, z_n]$ is the set of all infinite sequences of binary observations starting from $z_1, \dots, z_n$, and $z_1, z_2, \dots$ are the actual observations. The *lower benchmark* is

$$\mathrm{LB}_n := \frac{\mathrm{Markov}(\pi_{1|0}, \pi_{1|1})([z_1, \dots, z_n])}{\mathrm{Ber}(\hat\pi)([z_1, \dots, z_n])}, \tag{2}$$

where $\hat\pi := k/n$ (the maximum likelihood estimate) and $k = k(n)$ is the number of 1s among $z_1, \dots, z_n$. By definition, $\mathrm{UB}_0 = \mathrm{LB}_0 := 1$.

The upper benchmark (a likelihood ratio) is a martingale only under $\mathrm{Ber}(0.5)$ (and not under any other element of the null hypothesis), and so impossible to attain with "honest" methods such as conformal testing (since conformal test martingales are martingales under any element of the null hypothesis). The lower benchmark is valid under any element of the null hypothesis, but they do not generalize to complicated non-binary cases. Ramdas et al.'s (2022) R process is the integral of the lower benchmark (2) over all Markov alternatives w.r. to a specific probability measure on them, a Jeffreys-type prior (and so it would be unfair to compare its performance with the performance of our conformal test martingales, which are aimed at a specific Markov distribution as alternative hypothesis).

The trajectories of the upper and lower benchmarks in the large scenario are shown in Figure 2 in red and green; the figure also shows the trajectory the R process, and the other

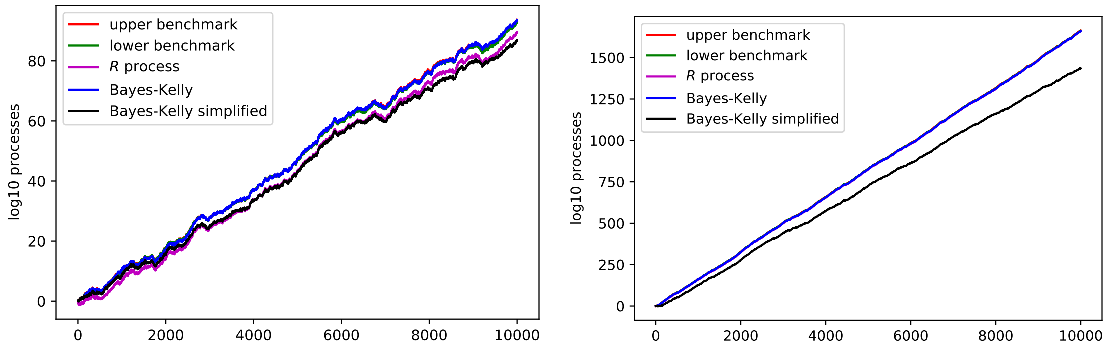

Figure 2: The two benchmarks, R process, Bayes–Kelly conformal test martingale, and its simplified version in the large scenario. The vertical axis is logarithmic (base 10). Left panel: hard case (the trajectories for the two benchmarks and Bayes–Kelly almost coincide). Right panel: easy case (the trajectories for the two benchmarks, R process, and Bayes–Kelly virtually coincide).

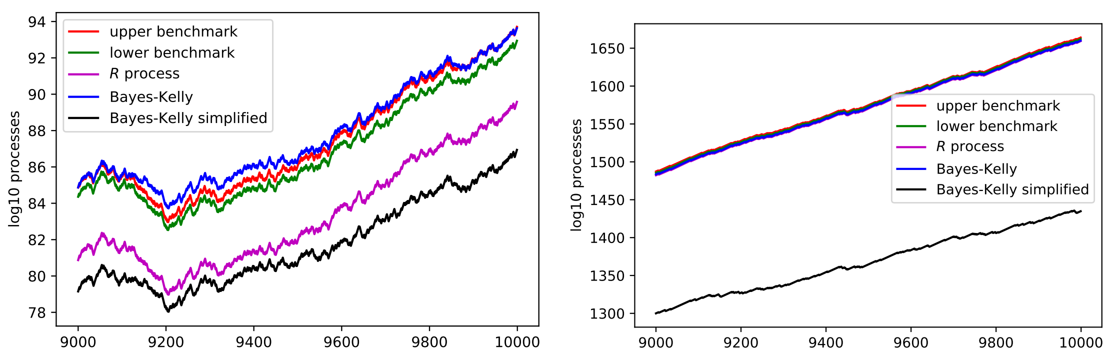

Figure 3: The analogue of Figure 2 for the last 1000 observations (in the right panel, the trajectories for the two benchmarks, R process, and Bayes–Kelly virtually coincide, as in Figure 2).

two trajectories should be ignored for now. The two benchmarks coincide or almost coincide. Figure 3 shows the same trajectories "under the lens", over the last 1000 observations. Notice that the upper benchmark can never be less than the lower benchmark.

4. Bayesian conformal testing

In this section we will use a Bayesian method that is statistically efficient (being competitive with the benchmarks) in our experiments but whose computational efficiency will be greatly improved in the next section. The p-values $p_1, p_2, \dots$ are generated as described in Vovk

(2021b); in particular, we are using the identity nonconformity measure (the nonconformity score of an observation z is z). Under the alternative hypothesis, the p-values are generated by a completely specified stochastic mechanism.

A conformal test martingale is usually represented as a product of betting functions, i.e., nonnegative functions $f : [0,1] \to [0,\infty)$ that integrate to 1. The value of the conformal test martingale after observing $z_1, \dots, z_n \in \{0,1\}$ is $f_1(p_1) \dots f_n(p_n)$, where $p_1, p_2, \dots$ are the conformal p-values and each betting function f_i is chosen, in a measurable manner, depending on $p_1, \dots, p_{i-1}$.

According to Fedorova et al. (2012, Theorem 2), the optimal (in the Kelly-type sense of that paper) betting functions f_n are given by the density of the predictive distribution of p_n conditional on knowing $p_1, \dots, p_{n-1}$. Let us find these predictive distributions. We will use the notation $U[a,b]$, where $a < b$, for the uniform probability distribution on the interval $[a,b]$ (so that its density is $1/(b-a)$).

We are in a typical situation of Bayesian statistics. The Bayesian parameter is the binary sequence $(z_1, z_2, \dots) \in \{0,1\}^\infty$ of observations, and the prior distribution on the parameter is Markov$(\pi_{1|0}, \pi_{1|1})$. The Bayesian observations are the conformal p-values $p_1, p_2, \dots$. Given the parameter, the distribution of p_n is

$$p_n \sim \begin{cases} U[0, k/n] & \text{if } z_n = 1 \\ U[k/n, 1] & \text{if } z_n = 0, \end{cases}$$

where $k := z_1 + \dots + z_n$ is the number of 1s among the first n observations.

Let $w^n_{k,L}$, where $n = 1, 2, \dots$, $k = 0, \dots, n$, and $L \in \{0,1\}$, be the total posterior probability of the parameter values $z_1, z_2, \dots$ for which $z_1 + \dots + z_n = k$ and $z_n = L$; we will use them as the weights when computing the predictive distributions for the p-values. We can compute the weights $w^n_{k,L}$ recursively in n as follows. We start from

$$w^1_{0,0} := w^1_{1,1} := 0.5, \quad w^1_{0,1} := w^1_{1,0} := 0. \tag{3}$$

At each step $n \geq 2$, first we compute the unnormalized weights

$$\tilde{w}^n_{k,0} := \left(w^{n-1}_{k,0} \pi_{0|0} + w^{n-1}_{k,1} \pi_{0|1} \right) l^{n-1}_k(0, p_n), \tag{4}$$

$$\tilde{w}^n_{k,1} := \left(w^{n-1}_{k-1,0} \pi_{1|0} + w^{n-1}_{k-1,1} \pi_{1|1} \right) l^{n-1}_{k-1}(1, p_n), \tag{5}$$

where l is the likelihood defined by

$$l^n_k(1,p) := \begin{cases} \frac{n+1}{k+1} & \text{if } p \leq \frac{k+1}{n+1} \\ 0 & \text{otherwise,} \end{cases}$$

$$l^n_k(0,p) := \begin{cases} \frac{n+1}{n-k+1} & \text{if } p \geq \frac{k}{n+1} \\ 0 & \text{otherwise,} \end{cases}$$

and then we normalize them:

$$w^n_{k,L} := \tilde{w}^n_{k,L} \Big/ \sum_{k=0}^{n} \sum_{L=0}^{1} \tilde{w}^n_{k,L}. \tag{6}$$

Algorithm 1: Bayes–Kelly $((p_1, p_2, \dots) \mapsto (S_1, S_2, \dots))$

$S_0 := S_1 := 1$
set the initial weights as per (3)
for $n = 2, 3, \dots$ **do**
  $S_n := f_n(p_n) S_{n-1}$, with f_n defined by (7)
  update the weights as per (4), (5), and (6)

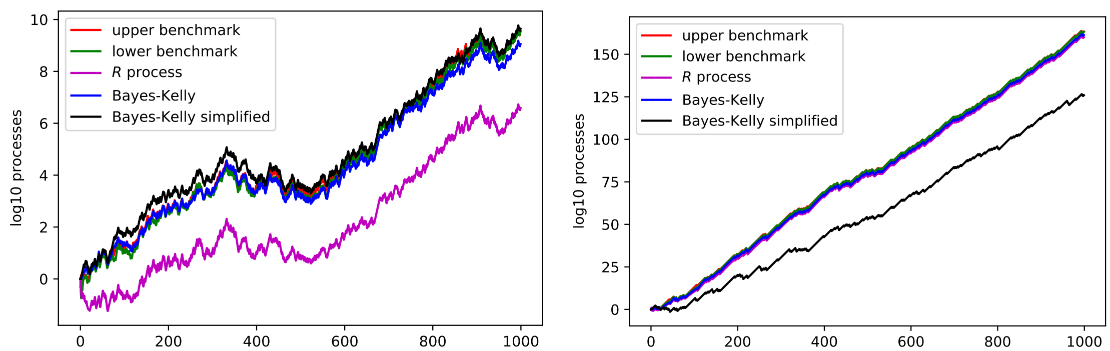

Figure 4: The Bayes–Kelly and Bayes–Kelly simplified conformal test martingales, the R-process, and the two benchmarks in the middle scenario. The vertical axis is logarithmic (base 10). Left panel: hard case. Right panel: easy case (the trajectories for the two benchmarks, R process, and Bayes–Kelly virtually coincide).

Given the posterior weights for the previous step, we can find the predictive distribution for p_n as

$$p_n \sim \sum_{k=0}^{n-1} \sum_{L=0}^{1} w_{k,L}^{n-1} \left(\pi_{1|L} U \left[0, \frac{k+1}{n} \right] + \pi_{0|L} U \left[\frac{k}{n}, 1 \right] \right),$$

where we use the shorthand $\pi_{0|L} := 1 - \pi_{1|L}$. Therefore, the betting functions for the resulting *Bayes–Kelly conformal test martingale* are

$$f_n(p) = \sum_{k=0}^{n-1} \sum_{L=0}^{1} w_{k,L}^{n-1} \left(\frac{n}{k+1} \pi_{1|L} 1_{p \le \frac{k+1}{n}} + \frac{n}{n-k} \pi_{0|L} 1_{p \ge \frac{k}{n}} \right). \tag{7}$$

The procedure is summarized as Algorithm 1; $S_0, S_1, \dots$ is the resulting trajectory of the Bayes–Kelly conformal test martingale.

For experimental results, see Figures 2 and 3 (large scenario) and Figure 4 (middle scenario). The Bayes–Kelly conformal test martingale appears to be very close to the two benchmarks in the middle scenario. Its simplified version is described in the next section. The relatively poor performance of the R process in the left panel of Figure 4 should not be interpreted as it being inferior to the Bayes–Kelly conformal test martingale: remember that R works against all Markov alternatives, whereas the other processes that we consider,

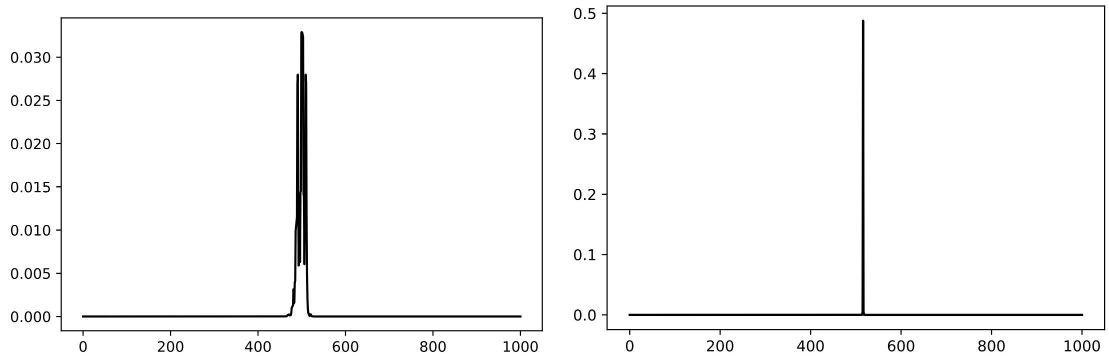

Figure 5: The weights w_k^{1000}, $k = 0, \dots, 1000$, at the last step for the Bayes–Kelly conformal test martingale in the medium scenario (the hard case on the left and easy on the right).

Algorithm 2: Simplified Bayes–Kelly $((p_1, p_2, \dots) \mapsto (S_1, S_2, \dots))$

$S_0 := S_1 := 1$
for $n = 2, 3, \dots$ **do**
  if $p_{n-1} \le 0.5$ **then** $L := 1$ **else** $L := 0$
  if $p_n \le 0.5$ **then** $S_n := 2\pi_{1|L} S_{n-1}$ **else** $S_n := 2\pi_{0|L} S_{n-1}$

including the two benchmarks, are adapted to a specific Markov alternative hypothesis (Markov$(0.4, 0.6)$ in the hard case and Markov$(0.1, 0.9)$ in the easy case). In Figure 3, the Bayes–Kelly conformal test martingale even exceeds the two benchmarks over some range of observations, but it is a statistical fluke.

5. Simplified Bayesian conformal testing

In this section we consider a radical simplification of the Bayes–Kelly conformal test martingale (7). We still assume that the Markov chain is symmetric, as in our model situations. (This assumption will be relaxed in the appendix.) If we assume that the weights $w_{k,L}^n$, $k = 0, \dots, n$, are concentrated at

$$k \approx k + 1 \approx n/2$$

and set

$$L := \begin{cases} 1 & \text{if } p_{n-1} \le 0.5 \\ 0 & \text{if not,} \end{cases} \tag{8}$$

(7) will simplify to

$$f_n(p) = 2\pi_{1|L} 1_{p \le 0.5} + 2\pi_{0|L} 1_{p > 0.5}, \tag{9}$$

with L defined by (8). Figure 5 shows the weights (averaged over $L \in \{0, 1\}$) for the last step of the Bayes–Kelly conformal test martingale in the medium scenario (10^3 observations).

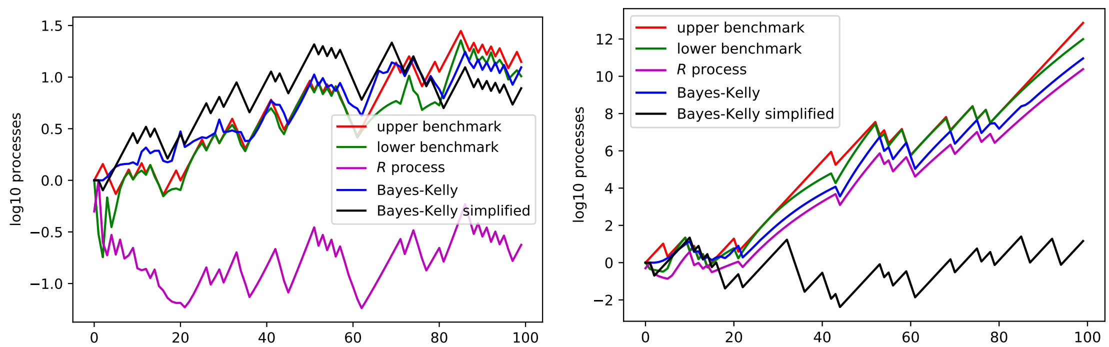

Figure 6: The analogue of Figures 2 and 4 for the small scenario (with the hard case on the left and easy on the right).

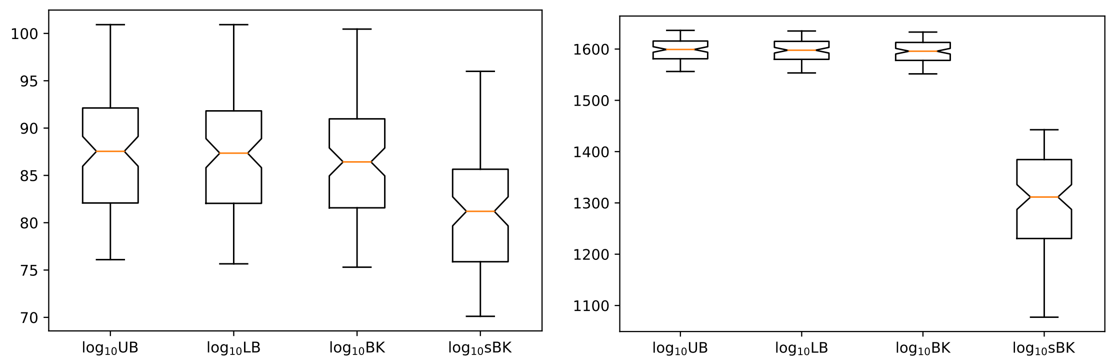

Figure 7: Boxplots based on 10^2 runs for the final values of the two benchmarks (upper UB and lower LB), the Bayes–Kelly conformal test martingale (BK), and its simplified version (sBK) in the large scenario. Left panel: hard case. Right panel: easy case.

They are indeed concentrated around values of k not so different from $0.5N = 500$; this is because $B_{0.5}$ is the stationary distribution in both hard and easy cases, both being symmetric. If $k(n-1) := z_1 + \cdots + z_{n-1} \approx (n-1)/2$, then the expression on the right-hand side of (8) is equal to z_{n-1} with high probability, which justifies setting (8). The simplified procedure is summarized as Algorithm 2. It is both simpler and more efficient computationally as compared with Algorithm 1.

The performance of Algorithm 2 is shown in Figures 2–4 and 6. It is usually worse than that of the Bayes–Kelly conformal test martingale and the two benchmarks, but is comparable on the log scale apart from the right panel of Figure 6.

The right panels of Figure 6 and Figures 8 and 9 show that the statistical performance of the simplified Bayes–Kelly martingale particularly suffers in the easy case. The notches in the boxplots in Figures 7–9 indicate confidence intervals for the median.

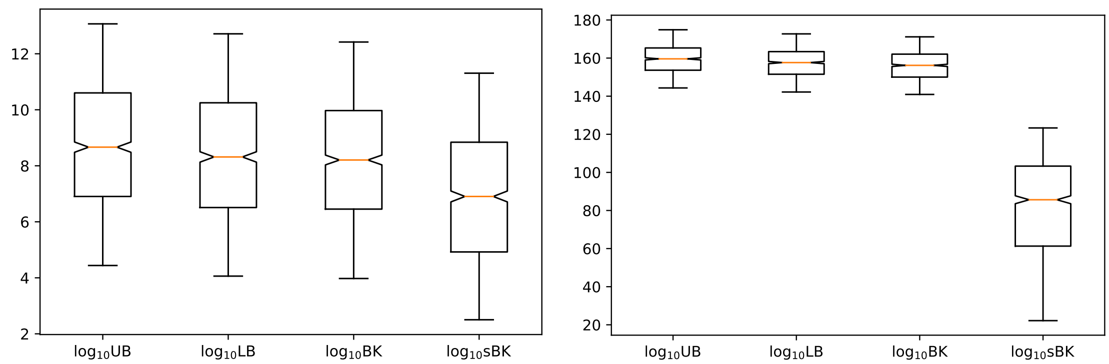

Figure 8: The analogue of Figure 7 for 10^3 runs in the medium scenario.

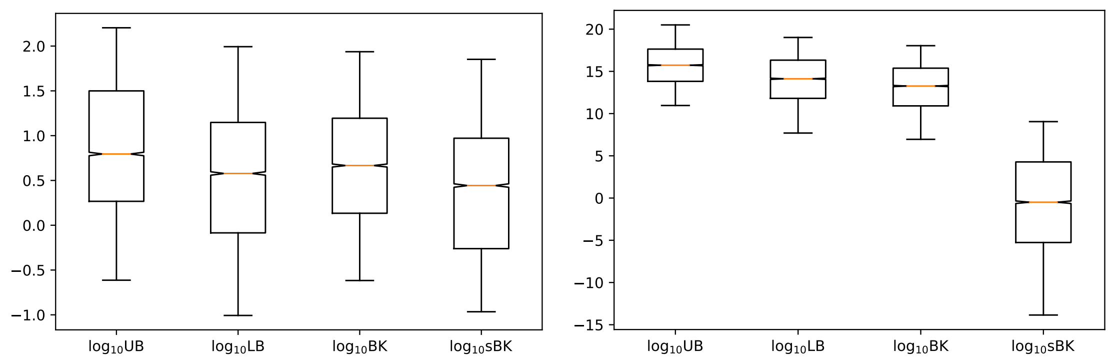

Figure 9: The analogue of Figures 7 and 8 for 10^4 runs in the small scenario.

6. Conclusion

This paper replicates, to a high degree of accuracy, the empirical results obtained by Ramdas et al. (2022) while using the method of conformal testing (based on conformal prediction). Ramdas et al.'s method is restricted to finite observation spaces, since their main object if study, referred to as the R process in this paper, is the lower benchmark (2) with the numerator replaced by another alternative distribution (namely, the average of all Markov distributions with respect to a Jeffreys-type prior). Even if the observation space is the real line $\mathbb{R}$ (let alone the sophisticated observation spaces used in machine learning), the lower benchmark becomes useless for interesting alternatives: the maximum likelihood estimate over all exchangeable distributions will become concentrated on the actual observations making the benchmark equal to zero (while identical 1 is achievable trivially by never gambling).

On the other hand, conformal testing is applicable in wide generality: see, e.g., Vovk (2021a), Vovk (2021b), and Vovk et al. (2022, Chapter 7). And its good performance (competitive with our benchmarks) in the binary case suggests that it is statistically efficient in general.

The most obvious direction of further research developing results of Vovk (2021b) and this paper is to try and adapt the Bayes–Kelly conformal test martingales (whether simplified or not) to infinite observation spaces, perhaps starting from the real line $\mathbb{R}$. A first step in this direction was made in Nouretdinov et al. (2021).

Acknowledgments

Many thanks to Martin Larsson and Johannes Ruf for useful discussions. We are grateful to Werner and David Carlisle for help with LaTeX. Comments by the three anonymous referees of the conference version of the paper helped us to improve the presentation.

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.

James R. Norris. *Markov Chains*. Cambridge University Press, Cambridge, 1997.

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

Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen. Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, 141:83–109, 2022.

Daniel Revuz and Marc Yor. *Continuous Martingales and Brownian Motion*. Springer, Berlin, third edition, 1999.

Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors, and p-values. *Statistical Science*, 26:84–101, 2011.

Albert N. Shiryaev. *Probability-2*. Springer, New York, third edition, 2019.

Vladimir Vovk. Testing randomness online. *Statistical Science*, 36:595–611, 2021a.

Vladimir Vovk. Conformal testing in a binary model situation. *Proceedings of Machine Learning Research*, 152:131–150, 2021b. COPA 2021.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, second edition, 2022. To appear.

Appendix A. Asymmetric Markov alternatives

In the main part of this paper we considered, following Ramdas et al. (2022, Section 4.2), the case of symmetric Markov alternatives (i.e., the case $\forall i,j : \pi_{i|j} = \pi_{j|i}$). In this appendix we do not assume symmetry and only assume $\min_{i,j} \pi_{i|j} > 0$; in particular, the Markov chain is aperiodic and irreducible. We still assume that the initial distribution of the Markov chain is uniform (although Proposition 2 below only needs the initial distribution to be positive, i.e., both probabilities, for 0 and for 1, to be positive).

The definition of the lower benchmark (2) still works in the asymmetric case, but in the definition of the upper benchmark (1) we replace $\text{Ber}(0.5)$ in the denominator by $\text{Ber}(\pi_1)$, where π_1 is the probability of 1 under the stationary distribution for the Markov chain. By definition, the stationary distribution (π_0, π_1), where π_0 is the probability of 0, satisfies

$$\begin{cases} \pi_{0|0}\pi_0 + \pi_{0|1}\pi_1 = \pi_0 \\ \pi_{1|0}\pi_0 + \pi_{1|1}\pi_1 = \pi_1. \end{cases} \tag{10}$$

By the ergodic theorem (Norris, 1997, Theorem 1.10.2), this choice of the denominator for the likelihood ratio process makes the upper benchmark as close to the lower benchmark as possible asymptotically. The following proposition says that this choice of the denominator is asymptotically optimal.

Proposition 2 *For any $x \in (0,1) \setminus \{\pi_1\}$,*

$$\frac{\text{Markov}(\pi_{1|0}, \pi_{1|1})([z_1, \dots, z_n])}{\text{Ber}(x)([z_1, \dots, z_n])} > \frac{\text{Markov}(\pi_{1|0}, \pi_{1|1})([z_1, \dots, z_n])}{\text{Ber}(\pi_1)([z_1, \dots, z_n])}$$

from some n on almost surely under $\text{Markov}(\pi_{1|0}, \pi_{1|1})$.

Proof We have, by the ergodic theorem and strong law of large numbers for martingales (Shiryaev, 2019, Theorem 7.5.4), almost surely as $n \to \infty$,

$$\begin{aligned}
&\frac{1}{n} \log \frac{\text{Markov}(\pi_{1|0}, \pi_{1|1})([z_1, \dots, z_n])}{\text{Ber}(x)([z_1, \dots, z_n])} \\
&= \pi_{0|0}\pi_0 \log \frac{\pi_{0|0}}{1-x} + \pi_{1|0}\pi_0 \log \frac{\pi_{1|0}}{x} + \pi_{0|1}\pi_1 \log \frac{\pi_{0|1}}{1-x} + \pi_{1|1}\pi_1 \log \frac{\pi_{1|1}}{x} + o(1) \\
&= \pi_{0|0}\pi_0 \log \frac{1}{1-x} + \pi_{1|0}\pi_0 \log \frac{1}{x} + \pi_{0|1}\pi_1 \log \frac{1}{1-x} + \pi_{1|1}\pi_1 \log \frac{1}{x} + c + o(1) \\
&= \pi_0 \log \frac{1}{1-x} + \pi_1 \log \frac{1}{x} + c + o(1) > \pi_0 \log \frac{1}{\pi_0} + \pi_1 \log \frac{1}{\pi_1} + c + o(1) \\
&= \frac{1}{n} \log \frac{\text{Markov}(\pi_{1|0}, \pi_{1|1})([z_1, \dots, z_n])}{\text{Ber}(\pi_1)([z_1, \dots, z_n])} + o(1),
\end{aligned}$$

where c is a constant (depending only on the πs), the penultimate "=" follows from (10), and the last inequality, ">", disregards the $o(1)$ terms and follows from the positivity of the Kullback–Leibler distance in this context. ■

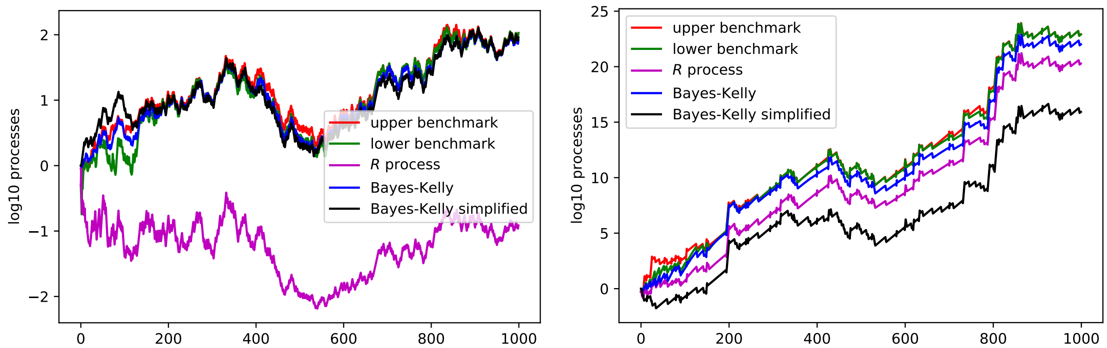

Figure 10: The analogue of Figures 2, 4, and 6 for the medium scenario and the asymmetric hard (left) and asymmetric easy (right) cases.

The Bayes–Kelly conformal test martingale (Algorithm 1) also works for asymmetric Markov chains. Let us derive the simplified Bayes–Kelly conformal test martingale (Algorithm 2) in the non-symmetric case. The solution to (10) is

$$\pi_1 = \frac{\pi_{1|0}}{\pi_{1|0} + \pi_{0|1}}.$$

When

$$k \approx k + 1 \approx n\pi_1$$

and

$$L := \begin{cases} 1 & \text{if } p_{n-1} \le \pi_1 \\ 0 & \text{if not,} \end{cases}$$

(7) will lead to

$$f_n(p) = \frac{\pi_{1|L}}{\pi_1} 1_{p \le \pi_1} + \frac{\pi_{0|L}}{\pi_0} 1_{p > \pi_1}$$

in place of (9).

Examples of the performance of various processes in simulation studies with asymmetric Markov alternatives are shown in Figures 10 and 11 (the poor performance of the R process in the left panel of Figure 10 should be ignored, since the comparison is not fair, as discussed earlier). In the *asymmetric hard case* the model is Markov$(0.4, 0.5)$, and in the *asymmetric easy case*, the model is Markov$(0.1, 0.5)$. (These two cases are somewhat harder than the symmetric hard and easy cases, respectively.)

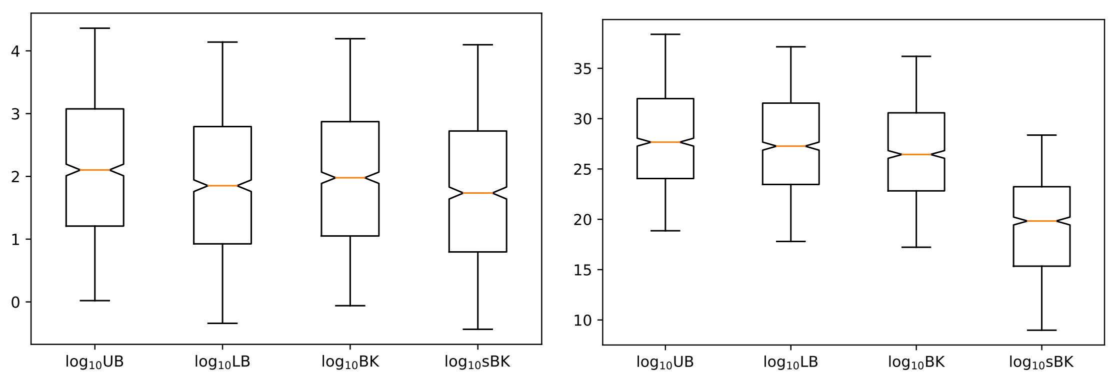

Figure 11: The analogue of Figures 8 and 9 for the medium scenario and the asymmetric hard (left) and asymmetric easy (right) cases.