An optimality property of the Bayes–Kelly algorithm

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 #39

First posted February 4, 2024. Last revised February 5, 2024.

Project web site:
http://alrw.net

Abstract

This note states a simple property of optimality of the Bayes–Kelly algorithm for conformal testing and poses a related open problem.

Contents

1 Introduction

This note is motivated by Grünwald et al.'s RSS discussion paper [1]. The main result of that paper (Theorem 1) is elegant and satisfying, but in this note I will concentrate on one of its limitations. The two modest contributions of my comment is to state a very simple property of optimality of the Bayes–Kelly algorithm [4, Sect. 9.2.1] in conformal testing and to pose an open problem. I will try to make the conformal testing part formally self-contained, but for further intuition and details, see [4].

A *test martingale* S in a probability space $(\Omega, \mathcal{F}, P)$ equipped with a filtration $(\mathcal{F}_n)_{n=0}^{\infty}$ is a nonnegative martingale starting from 1. In other words, $S = (S_0, S_1, \dots)$, each S_n is required to be $\mathcal{F}_n$-measurable, $\mathbb{E}_P(S_n \mid \mathcal{F}_{n-1}) = S_{n-1}$ for all $n \geq 1$, S_n are required to be nonnegative, and $S_0 = 1$. We can interpret $\log S_n$ as the amount of evidence (measured in bits) found at time n against P as null hypothesis. One-step counterparts of test martingales are "e-variables", to be introduced shortly.

2 Grünwald et al.'s result with discussion

The following is Grünwald et al.'s [1] main result (in its basic form, namely Theorem 1 in [1, Sect. 2]). The notation used in it will be explained after the statement.

Theorem 2.1 (Grünwald et al.)**.** *Suppose Q is a probability distribution with full support and with density q, and assume*

$$\inf_{w \in \mathcal{W}} D(Q \| P_w) < \infty. \tag{1}$$

Then there exists a (potentially sub-) distribution P with density p such that

$$E^* := q/p \tag{2}$$

is an e-variable. Moreover, E^* satisfies, essentially uniquely,

$$\sup_{E \in \mathcal{E}} \mathbb{E}_Q(\log E) = \mathbb{E}_Q(\log E^*) = \inf_{W \in \mathcal{W}} D(Q \| P_W) = D(Q \| P). \tag{3}$$

If the inf *is attained, so that $D(Q \| P_{W^*}) = D(Q \| P)$, then $P = P_{W^*}$.*

Theorem 2.1 is about a statistical model $(P_\theta \mid \theta \in \Theta)$ such that each P_θ is absolutely continuous w.r. to some underlying measure μ. This statistical model plays the role of our null hypothesis, while Q plays the role of the alternative hypothesis. The notation $D(Q \| P)$ refers to the Kullback–Leibler divergence between Q and P. The parameter space Θ is equipped with a σ-algebra, $\mathcal{W}$ stands for the set of all probability measures on Θ, and $\mathcal{E}$ stands for the set of all *e-variables*, i.e., nonnegative random variables E satisfying $\mathbb{E}_{P_\theta}(E) \leq 1$ for all $\theta \in \Theta$. For each $W \in \mathcal{W}$, $P_W := \int P_\theta W(\mathrm{d}\theta)$ is the mixture of P_θ w.r. to W. "Essentially uniquely" is defined in a natural way.

The first equality in (3) can be interpreted as saying that E^* is the optimal e-variable under Q.

2.1 Limitation

An important limitation of Theorem 2.1, as I see it, is the condition (1). This condition is stated in exactly this way only in Grünwald et al.'s Theorem 1 in its basic form of [1, Sect. 2], but analogous conditions are present in all generalizations given in [1].

To see how restrictive (1) is, consider, following [2] and [4, Chap. 9], the case of coin tossing. Formally, the null hypothesis is $(P_\theta \mid \theta \in [0,1])$, where $P_\theta = B_\theta^\infty$ and B_θ is the probability measure on $\{0,1\}$ satisfying $B_\theta(\{1\}) = \theta$. The condition (1) then means that Q should be absolutely continuous w.r. to an exchangeable probability measure on $\{0,1\}^\infty$. This condition is violated for interesting Q, such as a Markov Q outside the family $(P_\theta \mid \theta \in [0,1])$. The main alternative hypothesis used in [2] is a Jeffreys-type mixture of the Markov measures, and then (1) is also violated. It is difficult to think of natural cases where (1) would be satisfied.

2.2 An objection

A possible objection to the argument of Sect. 2.1 is that infinite sequences are irrelevant in real life, where we only observe finite sequences. If instead of the sample space $\{0,1\}^\infty$ we consider the sample space $\{0,1\}^n$ for a finite (even if very large) n, the condition (1) will be satisfied. Isn't it all that matters?

The difficulty with this solution is that the e-variables E_n^* obtained in this way by using Theorem 2.1 do not have to cohere with each other for different n; for example, they do not have to form a test martingale, or a test supermartingale, or an e-process. Optional continuation and stopping become big problems for E_n^*, and $\log E_n^*$ are no longer jointly valid as the amount of evidence found against the null hypothesis at time n.

To summarise, yes, we can truncate the sequential process of observing the bits $z_1, z_2, \dots \in \{0,1\}$, but it matters where we truncate it. On the other hand, if we build an e-variable $E^* : \{0,1\}^\infty \to [0,\infty]$ for the infinite time horizon, Lévy's "upward" theorem [6, Theorem 14.2] will give us a test martingale. Therefore, the kind of infinity that we need as our time horizon is the potential infinity, not the actual one. If we do not know the number of observations in advance and just would like to have an online measure of evidence found against the null hypothesis, Theorem 2.1 does not give us anything useful. Replacing $E^* : \{0,1\}^\infty \to [0,\infty]$ by $E^* : \{0,1\}^N \to [0,\infty]$ for a very large N is also awkward since the resulting test martingale will typically depend on N (even over the first few steps n).

2.3 Simple solution and its limitation

It is interesting that the binary case with a given alternative hypothesis (such as mixed Markov [2] or change-point [4, Sect. 9.2.3]) admits a simple solution:

just replace (2) by

$$E_n^* := \frac{Q([Z_1,\dots,Z_n])}{\sup_{\theta\in[0,1]} P_\theta([Z_1,\dots,Z_n])}$$

for each step $n = 0, 1, \dots$, where $[z_1,\dots,z_n]$ stands for the set of all infinite sequences in $\{0,1\}^\infty$ that begin with $z_1,\dots,z_n$, and Z_i are the random bits whose realizations are the observed bits z_i. This corresponds to replacing the mean p of the densities p_θ of P_θ (p is the mean if we assume that the inf is attained according to the last statement of Theorem 2.1) by the supremum of p_θ. Then E_n^* agree with each other in the sense of forming an e-process [2, Theorem 6].

However, the idea of replacing p in (2) by sup does not work outside narrow parametric cases [4, Remark 9.8]. The following example is still very basic (in machine learning the task is usually to predict the labels of complicated objects, such as movies).

Example 2.2. Fix a finite time horizon $N \gg 1$ and assume that the observations $z_1,\dots,z_N$ are real numbers, $z_n \in \mathbb{R}$, so that $Z_1,\dots,Z_N$ are random variables. The null hypothesis is that of *randomness*: $Z_1,\dots,Z_N$ are IID. The alternative hypothesis is a continuous probability measure Q on $\mathbb{R}^N$ (such as a changepoint hypothesis, as in [4, Remark 9.8]). Then the likelihood ratio of Q to the maximum likelihood over the null hypothesis is 0; indeed, $Q([z_1,\dots,z_N]) = 0$ and the maximum likelihood is positive, namely at least N^{-N} (it is exactly N^{-N} if the N observations are all different). This is worse than useless, as the identical 1 is a trivial test martingale.

3 Conformal testing

Our book [4, Part III] presents a general framework, which we call conformal testing, for testing nonparametric null hypotheses (first of all the hypothesis of randomness) in very general situations typical of machine learning. The procedure does not depend on assumptions such as (1). The efficiency of conformal testing is demonstrated in empirical studies reported in [4, Chap. 8], but theoretical results about efficiency have been established only in toy binary situations [4, Sect. 9.2]. The validity is, however, guaranteed, in that conformal testing leads to stochastic processes that are test martingales whenever the observations are IID.

Let me first introduce some terminology and notation. The *observation space* $\mathbf{Z}$ is a measurable space. (In the previous section we had $\mathbf{Z} = \{0,1\}$ and then, in Example 2.2, $\mathbf{Z} = \mathbb{R}$.) There is an underlying probability space $(\Omega, \mathcal{F}, P)$ in the background, but we rarely need it explicitly. We observe random elements $Z_1, Z_2, \dots$ of $\mathbf{Z}$ (formally, each Z_n is a measurable mapping from Ω to $\mathbf{Z}$) with their realized values denoted by $z_1, z_2, \dots \in \mathbf{Z}$.

The null hypothesis considered in this section (which we call the hypothesis of randomness) is that the observations $Z_1, Z_2, \dots$ are IID. (Generalization to other null hypotheses is briefly discussed in Sect. 4.) The underlying probability space is assumed to be rich enough; in particular, a probability measure on Ω making $Z_1, Z_2, \dots$ IID is assumed to exist (conformal prediction also needs a sequence of independent and uniformly distributed $\tau_1, \tau_2, \dots \in [0,1]$ modelling a random number generator).

We let $\mathbf{Z}^{(*)}$ stand for the set of all *bags* (or *multisets*) $\lbag z_1, \dots, z_n \rbag$ consisting of elements of $\mathbf{Z}$ (with $n = 0$ allowed); the difference between the bag $\lbag z_1, \dots, z_n \rbag$ and the set $\{z_1, \dots, z_n\}$ is that the bag (while still unordered) can contain several copies of the same element.

A conformal test martingale is determined by two components:

- A *conformity measure* A, which is a measurable function $A : \mathbf{Z}^{(*)} \times \mathbf{Z} \to \mathbb{R}$.
- A *betting martingale* B, which is a test martingale in the probability space $([0,1]^\infty, \mathcal{U}, U)$ with filtration $(\mathcal{U}_n)_{n=1}^\infty$, where $\mathcal{U}$ is the Borel σ-algebra on $[0,1]^\infty$, U is the uniform probability measure on $([0,1]^\infty, \mathcal{U})$, and $\mathcal{U}_n$ is the σ-algebra generated by the first n elements of the sequences in $[0,1]^\infty$.

Given these two components, we define the corresponding conformal test martingale as follows.

The *nth conformal p-value* p_n is defined by

$$p_n := \frac{|\{i : \alpha_i < \alpha_n\}| + \tau_n |\{i : \alpha_i = \alpha_n\}|}{n}, \tag{4}$$

where $i = 1, \dots, n$, the *conformity scores* α_i are computed from z_i using the conformity measure A by

$$\alpha_i := A(\lbag z_1, \dots, z_n \rbag, z_i), \quad i = 1, \dots, n, \tag{5}$$

and $\tau_1, \tau_2, \dots$ are independent random variables that are distributed uniformly on $[0,1]$ (modelling a random number generator). The *conformal test martingale* (*CTM*) S determined by A and B is the result of applying the betting martingale B to the p-values (4):

$$S_n := B_n(p_1, p_2, \dots), \quad n = 0, 1, \dots.$$

The associated σ-algebras $\mathcal{F}_n$ are those generated by $p_1, \dots, p_n$ (in particular, $\mathcal{F}_0 = \{\emptyset, \Omega\}$). An equivalent definition of a CTM is that it is a test martingale in the filtration $(\mathcal{F}_n)$ (this follows from, e.g., [6, Lemma A3.2]).

The property of validity for the conformal p-values is that they are independent and uniformly distributed on $[0,1]$ under the null hypothesis. This implies that a CTM is a test martingale under the null hypothesis.

Theoretical results about efficiency are established in [4, Chap. 9] only in the binary case and for a specific conformity measure (the identical one, $A(\lbag z_1, \dots, z_n \rbag, z_i) := z_i$). In this note we will take a slightly wider approach: will fix a conformity measure and will then find the optimal betting martingale for a given alternative hypothesis. Therefore, we consider a very limited kind of optimality:

- First, we restrict ourselves to the class of conformal test martingales.
- And even within this class, we consider a fixed conformity measure.

3.1 Full alternative hypotheses

Let Q be a probability measure on $\mathbf{Z}^\infty$; this is our alternative hypothesis about the distribution of $Z_1, Z_2, \ldots$. It is *full* in the sense of fully determining the distribution of the observations $Z_1, Z_2, \ldots$; in Sect. 3.2 we will consider an alternative probability measure on a poorer σ-algebra. Our null hypothesis, as before, is that $Z_1, Z_2, \ldots$ are IID.

The following paragraph is a description of the optimal (in the sense to be described later) under Q CTM with a given conformity measure A. Computationally efficient (or at least more explicit) versions of this CTM will be referred to as the *Bayes–Kelly algorithm*.

Consider the following Bayesian model (a statistical model plus a prior distribution on the parameter space). The parameter space is $\mathbf{Z}^\infty$, and it is equipped with Q as prior distribution. The element P_ζ of the statistical model indexed by $\zeta = (z_1, z_2 \ldots) \in \mathbf{Z}^\infty$ is the distribution of the corresponding p-values defined by (4) and (5) for a given ζ (we regard the random number generator $\tau_1, \tau_2, \ldots$ used in (4) as fixed). The marginal distribution of the p-values $p_1, p_2, \ldots$ is the mixture

$$P := \int P_\zeta Q(\,\mathrm{d}\zeta).$$

The relative increment S_n/S_{n-1} of the betting martingale on step n is then defined as the conditional density f_n of p_n given $p_1, \ldots, p_{n-1}$ (we will choose a natural version of the conditional density given by Lemma 3.1 below) evaluated at the realized p_n. Knowing S_n/S_{n-1} defines the betting martingale, since we know that its starting value is $S_0 = 1$. The *Bayes–Kelly CTM* is determined by A and this betting martingale.

Lemma 3.1. *A conditional density of p_n given $p_1, \ldots, p_{n-1}$ exists. There is a version f_n of the conditional density that is constant over each of the intervals*

$$\begin{aligned} &[i/n, (i+1)/n), \quad i = 0, \ldots, n-2, \\ &[(n-1)/n, 1]. \end{aligned} \tag{6}$$

Proof. For a fixed $\zeta \in \mathbf{Z}^\infty$, P_ζ generates independent p-values $p_1, p_2, \ldots$, and each p_n (defined by (4)) is distributed uniformly on an interval $[n_*/n, n^*/n]$ for some $n_* \in \{0, \ldots, n-1\}$ and $n^* \in \{i+1, \ldots, n\}$. Namely,

$$\begin{aligned} n_* &= |\{i \in \{1, \ldots, n\} : \alpha_i < \alpha_n\}| \\ n^* &= |\{i \in \{1, \ldots, n\} : \alpha_i \le \alpha_n\}| \end{aligned} \tag{7}$$

in the notation of (4). Let f_n^ζ be the density for p_n under Q conditional on knowing ζ and w.r. to the uniform probability measure on $[0,1]$. This is a piecewise constant function, which we assume taking constant values on the intervals (6). On each of these intervals, f_n^ζ takes values in $[0, n]$. Then f_n, as the integral of f_n^ζ w.r. to the posterior distribution on ζ (with prior Q and after observing $p_1, \ldots, p_{n-1}$), exists and integrates to 1, the latter following from Fubini's theorem. □

Algorithm 1 Bayes–Kelly algorithm (continuous version)

```
 1: $S_0 := 1$
 2: $\Sigma := \mathbf{Z}$
 3: for $n = 1, 2 \ldots$:
 4:     Set $f_n$ to the density of the pushforward of $Q_\Sigma$ under (4)–(5)
 5:     Read $z_n \in \mathbf{Z}$
 6:     Compute $\alpha_1, \ldots, \alpha_n$ as per (5)
 7:     Read $\tau_n \in [0, 1]$
 8:     Compute $p_n$ as per (4)
 9:     $S_n := S_{n-1} f_n(p_n)$
10:     for $(z_1, \ldots, z_n) \in \Sigma$:
11:         Compute $\alpha_1, \ldots, \alpha_n$ as per (5)
12:         Compute $n_*$ and $n^*$ as per (7)
13:         if $p_n \notin [n_*/n, n^*/n]$:
14:             Remove $(z_1, \ldots, z_n)$ from $\Sigma$
15:     Update $\Sigma := \Sigma \times \mathbf{Z}$
```

In [4, Chap. 9], we spell out the details of the Bayes–Kelly algorithm in two special (binary) cases:

- changepoint alternatives,
- Markov alternatives, following [2].

In both cases, the procedure is computationally efficient (while the computational efficiency of Algorithm 1 described below is unclear). We also prove its general optimality properties (i.e., without the *a priori* restriction to conformal testing).

Algorithm 1 is a version of the Bayes–Kelly algorithm that works for a conformity measure A that is continuous under the true data-generating distribution; in fact, it is sufficient to assume that, for all n and almost all $z_1, \ldots, z_n$, the n conformity scores (5) are all different.

Algorithm 1 maintains a set Σ of sequences $(z_1, \ldots, z_n)$ compatible with the p-values $p_1, \ldots, p_{n-1}$ observed so far; it is initialized to $\Sigma := \mathbf{Z}$ for $n = 1$ (line 2). The notation Q_Σ in line 4 stands for the conditional distribution of the first n observations generated from Q given that they belong to Σ. For $n = 1$ we have $f_1 := 1$. For this and other n, f_n is the pushforward of Q_Σ under the mapping of the type $\mathbf{Z}^n \to [0, 1]$ defined in two steps: first we apply (5) to the input $(z_1, \ldots, z_n)$ (ranging freely over $\mathbf{Z}^n$) and then we apply (4). Similarly, the variables $z_1, \ldots, z_n$ in the second **for** loop (starting in line 10) in Algorithm 1 are local ones; they range freely over $\mathbf{Z}$ and do not interfere with the global $z_1, \ldots, z_n$, which are the first n observations. Finally, the $\alpha_1, \ldots, \alpha_n$ inside that loop are local variables that are completely separate from the global variables with the same name. In line 12 we may assume $n^* = n_* + 1$ (this can be violated only with probability zero).

It should be clear how to drop the assumption of continuity of the conformity measure A in Algorithm 1: a sequence $z_1, z_2, \ldots$ leading to

$$k_n := |\{i = 1, \ldots, n : \alpha_i = \alpha_n\}| > 1$$

should have its weight multiplied by $1/k_n$ at step n after observing p_n that is compatible with this sequence.

The optimality property of the Bayes–Kelly algorithm is akin to Grünwald et al.'s one, namely to the first equality in (3). The next theorem will spell it out. In its statement, $\mathcal{S}_A$ is the class of all CTMs based on a conformity measure A, and $\mathcal{P}$ (representing the null hypothesis of randomness) consists of all P^∞, P ranging over the probability measures on $\mathbf{Z}$.

Theorem 3.2. *Fix a conformity measure A and an alternative hypothesis Q. At each step N, the Bayes–Kelly CTM S^* attains the maximum of $\mathbb{E}(\log S_N)$ among all CTMs S based on A:*

$$\sup_{S \in \mathcal{S}_A} \mathbb{E}_Q(\log S_N) = \mathbb{E}_Q(\log S^*_N) = D(A_* Q \| A_* P), \tag{8}$$

where $A_* Q$ is the pushforward of Q under the mapping (4)–(5) of generating the p-values $p_1, \ldots, p_N$, P is an arbitrary element of $\mathcal{P}$, and $A_* P$ (which is the uniform distribution on $[0,1]^N$) is defined analogously to $A_* Q$.

Notice the similarity between (3) and (8) (and we can also add $\inf_{P \in \mathcal{P}}$ in front of $D(A_* Q \| A_* P)$ in (8)). The property of optimality given in Theorem 3.2 is typical of Bayesian algorithms.

Proof of Theorem 3.2. The optimization problem $\mathbb{E}_Q(\log S_N) \to \max$ decomposes into the sequence of problems $\mathbb{E}_Q(\log(S_n/S_{n-1})) \to \max$ for $n = 1, \ldots, N$ and for given $p_1, \ldots, p_{n-1}$. It suffices to apply [4, Lemma 9.6]. □

Unlike Theorem 2.1, Theorem 3.2 gives e-variables S^*_n that are coherent in the sense of forming a test martingale. The reason for this is that using conformal p-values reduces the massive null hypothesis of randomness to the simple hypothesis of uniformity.

3.2 Shrunk alternative hypotheses

This subsection will be very speculative.

The Bayes–Kelley algorithm does not cover some interesting cases. The conformity measures A can be very complex, and then the problem of designing an optimal betting martingale becomes infeasible. It is even possible for A to have an element of intelligence in it [4, Sect. 8.6.3]. For example, A can be based on a deep neural network as an underlying algorithm [4, Sect. 4.3.3]. In this case the Bayes–Kelly algorithm will be infeasible even if we fix an alternative probability measure Q on $\mathbf{Z}^\infty$. (And we are unlikely to have such a probability measure Q in the first place if we are contemplating the use of such a conformity measure A.)

In the case of such a "quasi-intelligent" conformity measure, to find a suitable betting martingale B, we might consider an alternative distribution on the p-values $p_1, p_2, \dots$ (whose distribution is uniform under the null hypothesis) rather than an alternative distribution on the observations or (which is not very different) on the underlying probability space. Let us call a probability measure on $[0,1]^\infty$ interpreted as alternative distribution on the p-values a *shrunk alternative* (in the spirit of "filtration shrinkage", a popular topic of research in probability theory). The betting martingale B that is optimal in a natural sense will then be the likelihood ratio of the shrunk alternative to the null hypothesis of the uniform distribution on $[0,1]^\infty$.

How do we choose the shrunk alternative Q? A natural approach is simply to try and make it as large as possible in an attempt to approximate the universal distribution in the sense of the algorithmic theory of randomness (see, e.g., [5, Sect. 5], which calls it "*a priori* semidistribution", or [3, Appendix A], which considers a non-sequential setting). Such a betting martingale may be called "quasi-universal"; see [3, Appendix B] for a further discussion in a non-sequential setting. The quasi-universal betting martingale is designed to approximate the universal supermartingale (see, e.g., [5, Sect. 3]).

To summarise, an appealing informal choice is:

- a quasi-intelligent conformity measure;
- a quasi-universal betting martingale.

To make testing methods based on these choices computationally efficient, we might need modifications, e.g., in the direction of inductive conformal prediction [4, Sect. 4.2].

When designing a quasi-intelligent conformity measure, we still need an objective function, perhaps informal. In conformal prediction [4, Part I] a typical informal objective function is the conformity measure's sensitivity to unusual observations (and so detecting their "nonconformity"), which leads to smaller p-values for non-IID data. Motivated by this informal objective function, in the first edition of [4] (Sect. 7.1 of the first edition) we only considered betting martingales S such that, for each n, S_n is a decreasing function of the nth p-value p_n. However, later [4, Sect. 8.6.1] it turned out that allowing non-decreasing S_n greatly improves the performance of conformal test martingales on benchmark datasets. Now it appears that our informal objective function in designing a good conformity measure should be not to minimize the p-value (as it usually is in conformal prediction [4, Sect. 3.1]) but to come up with the most informative conformity scores α_i capturing the most relevant features of z_i.

4 Conclusion

In this note we concentrated on the case of the hypothesis of randomness as the null. In fact conformal testing is applicable to many other "online compression models"; see [4, Part IV]. Examples include partial exchangeability, Gaussian, hypergraphical (useful for causal inference), and Markov models.

The most obvious open problem arising in connection with Theorem 3.2 is whether there is a way of extending the Bayes–Kelly algorithm and (8) to the case when the conformity measure A also needs to be chosen optimally. When only given the alternative hypothesis Q, how do we choose the pair (A, B), where B is the betting martingale, optimally? In this note we have only discussed how to choose B optimally given Q and A.

Of course, it is not surprising that Theorem 2.1 of [1] has limitations (and the authors of [1] discuss some), but it is a great first step, and it opens up interesting directions of further research.

Acknowledgements

I am grateful to attendees of the RSS meeting where [1] was presented, in particular Peter Grünwald, Wouter Koolen, and Johannes Ruf, for useful discussions.

References

[1] Peter Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing. Technical Report arXiv:1906.07801 [math.ST], arXiv.org e-Print archive, March 2023. Journal version is to appear in the *Journal of the Royal Statistical Society B* (with discussion).

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

[3] Vladimir Vovk. Testing exchangeability in the batch mode with e-values and Markov alternatives. Technical Report arXiv:2305.05284 [stat.ME], arXiv.org e-Print archive, May 2023.

[4] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, Cham, second edition, 2022.

[5] Vladimir Vovk and Vladimir V. V'yugin. Prequential level of impossibility with some applications. *Journal of the Royal Statistical Society B*, 56:115–123, 1994.

[6] David Williams. *Probability with Martingales*. Cambridge University Press, Cambridge, 1991.