The power of forgetting

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 #37

April 13, 2023

Project web site:
http://alrw.net

How can forgetfulness and efficiency coexist? Aren't these two concepts absolute opposites? Far from it.

Mike Byster, "The Power of Forgetting", 2014

Abstract

This paper places conformal testing in a general framework of statistical hypothesis testing. A standard approach to testing a composite null hypothesis H is to test each of its elements and to reject H when each of its elements is rejected. It turns out that we can fully cover conformal testing using this approach only if we allow forgetting some of the data. However, we will see that the standard approach covers conformal testing in the long run, albeit under restrictive assumptions. I will also list several possible directions of further research, including developing a general scheme of online testing.

Contents

1 Introduction

Conformal testing is an interesting application of conformal prediction. It turns the p-values output by conformal prediction into dynamic procedures for testing statistical hypotheses. This paper is a high-level discussion of dynamic testing of statistical hypotheses, in an attempt to place conformal testing in the general theory of statistical hypothesis testing.

In conformal testing, we are usually interested in testing the exchangeability model, because of its importance in machine learning. In this paper, however, we will often consider testing other statistical models, which will shed new light on the relation of conformal testing to alternative approaches.

Section 2 briefly reviews the history of the online approach to hypothesis testing. Formal exposition starts in Sect. 3, which is a summary of three approaches to online testing, including, in Sect. 3.3, a summary of conformal testing. Section 4 introduces a general scheme covering all three approaches. Section 5 points out an unnatural feature of conformal testing, and Sect. 6 explains that the extent of the unnaturalness is limited (albeit under strong assumptions). Section 7 illustrates some of the points discussed in the earlier sections using computer simulations, and Sect. 8 concludes.

2 From batch to online hypothesis testing

The classical theory of statistical hypothesis testing, as created by Student (1908), Fisher (1925a), Egon Pearson, and Neyman (1933), was developed in the batch setting (in the terminology of modern machine learning). Given a batch of data $z_1, \dots, z_N$, we would like to test the hypothesis (known as the *null hypothesis*) that $z_1, \dots, z_N$ were generated from a given probability measure (in which case the null hypothesis is called *simple*) or a probability measure from a given family of probability measures (in which case the null hypothesis is called *composite*). The number of observations N (sample size) is chosen in advance. The classical theory is still dominant in statistical hypothesis testing.

Remark 2.1. I do not list Karl Pearson because he was interested in statistical tests, such as his famous χ^2 test [Pearson, 1900], that only have an asymptotic ("large-sample") justification. The first exact ("small-sample") test for an interesting composite null hypothesis was developed by Student (1908), whose results were rigorously proved and greatly developed by Fisher (1925a, 1925b, 1935).

The assumption that N is chosen in advance was removed during World War II by Wald [1945, 1947] in the US, with research along similar lines going on in the UK [Barnard, 1946]. However, Wald's picture was not fully dynamic: he just made N a stopping time when the decision (rejection or acceptance of the null hypothesis) is announced. The dynamic interpretation in which the likelihood ratio is interpreted directly as the evidence in favour/against the null/alternative hypothesis was given by Barnard [1947, pp. 459–460 and the last paragraph]. More recently, this interpretation has been widely discussed under the name of

the *law of likelihood* [Bandyopadhyay and Forster, 2011]. (The term "law of likelihood" was coined by Hacking 1965, Chap. 5, but he was only interested in its special case, namely in comparison of the likelihood ratio with 1.)

The dynamic way of testing a simple null hypothesis has its origin in Ville's (1939) notion of a martingale. The value of a test martingale (i.e., a nonnegative martingale with initial value 1) can be interpreted as the amount of evidence found against the null hypothesis. Ville did not have this interpretation in his book (infinite sequences were his main object of interest), but it formed gradually in the algorithmic theory of randomness; e.g., it is stated explicitly in Vovk and V'yugin [1994]. This interpretation is the basis of Shafer and Vovk [2001, 2019]. It is closely related to Barnard's (1947) mentioned earlier, since a test martingale can often be represented in the form of a likelihood ratio.

Remark 2.2. In particular, for simple null hypotheses, a test martingale is a likelihood ratio. Therefore, it has a very convincing Bayesian interpretation: if *a priori* we regard the null and the alternative (the numerator of the likelihood ratio) as equally probable, the posterior probability of the null will be $1/(L+1)$, where L is the likelihood ratio.

How do test martingales work for composite hypotheses? This is a standard question in the algorithmic theory of randomness: we test against each element of the composite null and then take the infimum. See, e.g., Vovk [1986], Vovk and V'yugin [1993, Theorem 2], Bienvenu et al. [2011, Sect. 4], and Gács [2021, Theorem 4.2.1].

This suggests gambling against all values of the parameter θ obtaining a test martingale S^θ for each θ and then taking the inf over θ. We will do this in Sect. 5.

3 Three modern ways of dynamic hypothesis testing

In this section we will discuss three approaches, by now standard, to dynamic hypothesis testing. Only one of them, conformal testing, can be used for testing the exchangeability model (the standard statistical model in machine learning) in non-trivial cases.

First we introduce our framework and notation. Let $(\Omega, \mathcal{F})$ be a measurable space equipped with a family P_θ, $\theta \in \Theta$, of probability measures on $(\Omega, \mathcal{F})$. We refer to $(\Omega, \mathcal{F})$ as our *sample space* and to $(P_\theta \mid \theta \in \Theta)$ as our *statistical model*. We are not assuming that the model is parametric (i.e., that Θ is a subset of a finite-dimensional Euclidean space $\mathbb{R}^n$); e.g., $(P_\theta \mid \theta \in \Theta)$ may be the set of all exchangeable probability measures on $\mathbb{R}^\infty$.

Our random observations are $Z_1, Z_2, \ldots$; these are random elements on $(\Omega, \mathcal{F})$ taking values in a measurable space $\mathbf{Z}$, which is our *observation space*. Set $\mathcal{F}_n := \sigma(Z_1, \ldots, Z_n)$ for $n = 0, 1, \ldots$, i.e., $\mathcal{F}_n$ is the σ-algebra generated by the first n observations. The interpretation of $\mathcal{F}_n$ is the information available

by time n. The sequence $(\mathcal{F}_n)$ of σ-algebras is called the *natural filtration*. Let $z_1, z_2, \dots$ be realizations of $Z_1, Z_2, \dots$.

Remark 3.1. It is more customary to start from the filtration $(\mathcal{F}_n)$ and require that each Z_n be measurable with respect to $\mathcal{F}_n$ for each n. This is a weaker requirement allowing further sources of information apart from the observations $Z_1, Z_2, \dots$. We will, however, assume that the observations are the only source of information, and will make similar assumptions in what follows.

We regard $(P_\theta \mid \theta \in \Theta)$ as our null hypothesis, and we would like to test whether $z_1, z_2, \dots$ were really generated from one of the P_θ.

If the statistical model contains only one probability measure P, online testing consists in choosing a *test martingale* S_n, $n = 0, 1, \dots$, i.e., a sequence of random variables such that S_n is $\mathcal{F}_n$-measurable, $S_0 = 1$, and, for each $n = 0, 1, \dots$,

$$\mathbb{E}(S_{n+1} \mid \mathcal{F}_n) = S_n.$$

We regard S_n as the capital at time n of a gambler betting against the null hypothesis P. Next we will discuss three known ways of generalizing this definition to composite null hypotheses.

3.1 Element-wise testing

The most basic and standard generalization is to gamble against each P_θ separately and to regard the null hypothesis falsified to the degree that all of P_θ have been falsified. Formally, for each $\theta \in \Theta$, we fix a test martingale S^θ, and we then define

$$S_n := \inf_{\theta \in \Theta} S^\theta_n. \tag{1}$$

Any process S that can be obtained in this way will be referred to as an *element-wise test*, and I will sometimes refer to this procedure of testing as *element-wise testing*.

Remark 3.2. The function $S^\theta_n(\omega)$ of $\theta \in \Theta$, n, and $\omega \in \Omega$ is not assumed to be measurable in θ, and so S_n is not a random variable in general. (And even if $S^\theta_n(\omega)$ were assumed measurable in θ, taking an inf over an uncountable set may destroy measurability.)

A special case of element-wise testing (1) is used in Ramdas et al. [2022], where each S^θ is defined as the likelihood ratio $\mathrm{d}Q/\mathrm{d}P_\theta$, where Q is a probability measure that does not depend on θ (while dependence on θ is allowed in element-wise testing in general). We will refer to this special case as *simple element-wise testing*.

Remark 3.3. Ramdas et al. [2022] apply their simple element-wise testing scheme to testing exchangeability, but, as we explain in Vovk et al. [2022, Sect. 9.2.1] (see, especially, Remarks 9.7 and 9.8), this scheme (based on the maximum likelihood estimate) is applicable to testing exchangeability only in toy situations.

3.2 Pivotal testing

The second approach goes back to Fisher's fiducial statistics and was widely promoted by, e.g., George Barnard and Donald Fraser. Recent work includes Peter McCullagh's (see McCullagh et al. 2009) and the work on confidence distributions, including confidence predictive distributions [Schweder and Hjort, 2016, Shen et al., 2018].

An *online pivotal model* is a pair (N, Q), where N is a measurable mapping (*normalizing transformation*) $N : \mathbf{Z}^* \to \mathbf{Z}'$ for some measurable space $\mathbf{Z}'$ and Q is a probability measure on $(\mathbf{Z}')^\infty$. We say that it *agrees* with our statistical model $(P_\theta \mid \theta \in \Theta)$ if the distribution of the random sequence

$$(Z'_1, Z'_2, \dots) := (N(Z_1), N(Z_1, Z_2), N(Z_1, Z_2, Z_3), \dots), \tag{2}$$

where $(Z_1, Z_2, \dots) \sim P_\theta$, is Q (in particular, it does not depend on the parameter θ). We might say that it *strongly agrees* with our statistical model if $\{P_\theta \mid \theta \in \Theta\}$ contains all probability measures on $(\Omega, \mathcal{F})$ for which (2) is distributed as Q; however, we will not use this stronger notion.

Remark 3.4. The definition of an online pivotal model can be trivially extended by allowing N to depend on the parameter value $\theta \in \Theta$. Such an extension would even better agree with the word "pivot", since in statistics pivotal quantities are allowed to depend on θ (those that do not depend on θ are usually called "ancillary statistics", but a disadvantage of the term "ancillary statistic" is that it is usually associated with conditional inference). In this paper we will only be interested in examples where the normalizing transformation N does not depend on θ.

Let me give three very simple examples of online pivotal models. The *full Gaussian pivotal model* is (N, Q) where

$$\begin{aligned} N(z_1) &:= 0, \\ N(z_1, \dots, z_n) &:= (z_n - z_1)/(z_2 - z_1) \quad \text{for } n \geq 2, \end{aligned}$$

and Q is the push-forward of $\mathcal{N}_{0,1}$ under N. (Let us set, e.g., $0/0 := 0$, here and below.) This online pivotal model agrees with the 2-parameter Gaussian statistical model $(\mathcal{N}_{\mu,\sigma^2} \mid \mu \in \mathbb{R},\ \sigma > 0)$ (which is parametrized by the mean μ and variance σ^2). The other two example are, in some sense, submodels of this model.

The *Gaussian pivotal model with variance 1* is (N, Q) where

$$N(z_1, \dots, z_n) := z_n - z_1 \quad \text{for } n \geq 1, \tag{3}$$

and Q is the push-forward of $\mathcal{N}_{0,1}$ under N. This online pivotal model agrees with the 1-parameter Gaussian statistical model $(\mathcal{N}_{\mu,1} \mid \mu \in \mathbb{R})$ with the variance fixed to 1. The *Gaussian pivotal model with mean 0* is (N, Q) where

$$N(z_1, \dots, z_n) := z_n / z_1 \quad \text{for } n \geq 1,$$

and Q is the push-forward of $\mathcal{N}_{0,1}$ under N. This online pivotal model agrees with another 1-parameter Gaussian statistical model, $(\mathcal{N}_{0,\sigma^2} \mid \sigma > 0)$ with the mean fixed to 0.

For further examples, see McCullagh et al. [2009] (Gauss linear model) and Ramdas et al. [2023, Sect. 4.1].

An online pivotal model reduces (perhaps not perfectly) a composite null model to a simple one, and gambling against a simple null model is unproblematic. Formally, set $\mathcal{F}'_n := \sigma(Z'_1, \dots, Z'_n)$ for $n = 0, 1, \dots$, and choose a test martingale S with respect to the filtration $(\mathcal{F}'_n)$ and probability measure Q. We will then refer to S as a *pivotal test martingale*.

Standard uses of online pivotal models are for producing prediction sets [McCullagh et al., 2009], confidence predictive distributions (Schweder and Hjort, 2016, Sect. 12.4; Shen et al., 2018), and confidence distributions (Cox, 1958; Xie and Singh, 2013; Schweder and Hjort, 2016). However, their adaptation to testing is straightforward, and is analogous to the step from conformal prediction to conformal testing.

3.3 Conformal testing

My exposition in this paper is intended to be self-contained (apart from the definition of BK test martingales in Sect. 7), but for further details about online compression models, see Vovk et al. [2022, Part IV].

An *online compression model* is a quadruple $(\Sigma, \square, F, B)$, where

- Σ is a measurable space, which is called the *summary space* and whose elements are called *summaries*;
- $\square \in \Sigma$ is a fixed summary called the *empty summary*;
- $F : \Sigma \times \mathbf{Z} \to \Sigma$ is a measurable function called the *forward function*;
- B is a Markov kernel mapping Σ to the probability measures on $\Sigma \times \mathbf{Z}$ such that
$$B(F^{-1}(\sigma \mid \sigma) = 1$$
for each $\sigma \in F(\Sigma \times \mathbf{Z})$.

An alternative, often more convenient (especially for defining specific examples) representation of online compression models is in terms of the corresponding repetitive structures. Namely, the *repetitive structure* corresponding to an online compression model $(\Sigma, \square, F, B)$ consists of the *summarising statistic* $t : \mathbf{Z}^* \to \Sigma$ defined by

$$\begin{aligned} t() &:= \square, \\ t(z_1, \dots, z_n) &:= F(t(z_1, \dots, z_{n-1}), z_n) \quad n = 1, 2, \dots, \end{aligned}$$

and of the inverse transformation mapping each $\sigma \in t(\mathbf{Z}^n)$ for each $n \in \{1, 2, \dots\}$ to the probability measure $P_n(\sigma)$ on $\mathbf{Z}^n$ defined by

$$P_n(\mathrm{d}z_1, \dots, \mathrm{d}z_n \mid \sigma_n) := B(\mathrm{d}\sigma_0, \mathrm{d}z_1 \mid \sigma_1) B(\mathrm{d}\sigma_1, \mathrm{d}z_2 \mid \sigma_2) \dots \\ B(\mathrm{d}\sigma_{n-2}, \mathrm{d}z_{n-1} \mid \sigma_{n-1}) B(\mathrm{d}\sigma_{n-1}, \mathrm{d}z_n \mid \sigma_n).$$

We say that a probability measure P *agrees* with the online compression model if, for each n, P_n is a version of the conditional probability, under P, of the first n observations given their summary. And we say that a statistical model $(P_\theta \mid \Theta)$ *agrees* with the online compression model if each P_θ does. (As in the case of online pivotal models, we do not require that $(P_\theta \mid \Theta)$ contain every probability measure that agrees with the online compression model.)

A *conformity measure* in an online compression model $(\Sigma, \square, F, B)$ is a measurable function $A : \Sigma \times \mathbf{Z} \to \overline{R}$. The *p-value* generated by the corresponding conformal predictor after observing $(z_1, \dots, z_n) \in \mathbf{Z}^n$ is

$$p_n := B_{\mathbf{Z}} \left(\{ z \in \mathbf{Z} \mid A(\sigma_n, z) < A(\sigma_n, z_n) \} \mid \sigma_n \right) \\ + \tau_n B_{\mathbf{Z}} \left(\{ z \in \mathbf{Z} \mid A(\sigma_n, z) = A(\sigma_n, z_n) \} \mid \sigma_n \right) \quad (4)$$

where $B_{\mathbf{Z}}$ is the marginal distribution

$$B_{\mathbf{Z}}(E \mid \sigma) := B(\Sigma \times E \mid \sigma)$$

and $\tau_n \in [0, 1]$ (in applications, a number produced by a random number generator).

The main property of validity of conformal prediction is that the p-values $p_1, p_2, \dots$ output according to (4) are independent and distributed uniformly on $[0, 1]$ provided the observations are generated from a probability measure that agrees with the online compression model and the random numbers $\tau_1, \tau_2, \dots$ are distributed uniformly on $[0, 1]$ and independent of the observations and between themselves. Let $\mathcal{F}'_n$ be the σ-algebra generated by $p_1, \dots, p_n$. A *conformal test martingale* is a test martingale with respect to the filtration $(\mathcal{F}'_n)$ and the uniform probability measure on $(p_1, p_2, \dots) \in [0, 1]^\infty$ (the latter determining the probability measure on $\sigma(\cup_n \mathcal{F}'_n)$ underlying the martingale).

Now we can give four standard examples of online compression models, which we do in terms of the corresponding repetitive structures. The *exchangeability model* has $\sigma_n = \wr z_1, \dots, z_n \int$ as the summary of a data sequence $(z_1, \dots, z_n)$, and $P_n(\sigma_n)$ is the uniform distribution on all orderings of σ_n (a fuller definition, dealing carefully with the possibility of repetitions among the element of σ_n, is that $P_n(\sigma_n)$ is the push-forward of the uniform probability measure on the $n!$ permutations $\pi : \{1, \dots, n\} \to \{1, \dots, n\}$ under the mapping $\pi \mapsto (z_{\pi(1)}, \dots, z_{\pi(n)})$). In fact, in all but one of our four examples, $P_n(\sigma_n)$ will be the uniform probability measure on $t_n^{-1}(\sigma_n)$, where t_n is the restriction of t to $\mathbf{Z}^n$.

In the remaining three examples, the observation space is the real numbers, $\mathbf{Z} := \mathbb{R}$. The *full Gaussian compression model* has the summarizing statistic

$$\sigma_n = t_n(z_1, \dots, z_n) := \left(\sum_{i=1}^n z_i, \sum_{i=1}^n z_i^2 \right)$$

(equivalently, the summary of a data sequence $z_1, \ldots, z_n$ consists of its empirical mean and standard deviation). The summarizing statistic for the *Gaussian compression model with variance 1* is

$$\sigma_n = t_n(z_1, \ldots, z_n) := \sum_{i=1}^{n} z_i, \tag{5}$$

and for the *Gaussian compression model with mean 0* it is

$$\sigma_n = t_n(z_1, \ldots, z_n) := \sum_{i=1}^{n} z_i^2.$$

The conditional distribution $P_n(\sigma_n)$ on $t_n^{-1}(\sigma_n)$ is defined to be the uniform distribution in the case of the full Gaussian compression model and Gaussian compression model with mean 0; in both of these cases $t_n^{-1}(\sigma_n)$ is a sphere, and the notion of the uniform distribution is meaningful and unambiguous. For the Gaussian model with variance 1, whose summarizing statistic is given by (5), $t_n^{-1}(\sigma_n)$ is not compact for $n > 1$, and the uniform distribution on it does not even exist; we define $P_n(\sigma_n)$ as the probability measure on $t_n^{-1}(\sigma_n)$ with density proportional to

$$\exp\left(-\frac{1}{2}\sum_{i=1}^{n} z_i^2\right). \tag{6}$$

Remark 3.5. The full Gaussian compression model (usually referred to simply as the Gaussian compression model) is the most general of our three Gaussian compression models, but it is easy to extend to a standard model of linear regression, the Gauss linear model, both in the pivotal [McCullagh et al., 2009] and conformal [Vovk et al., 2022, Sect. 11.4.2] cases.

It is interesting that the p-values $p_3, p_4, \ldots$ output by any of these three Gaussian compression models are almost surely deterministic (do not depend on the random numbers τ), while p_1 has the uniform distribution on $[0, 1]$. The second p-value p_2 behaves like p_1 in the case of the full Gaussian compression model and like $p_3, p_4, \ldots$ for the other two models.

Each Gaussian probability measure $\mathcal{N}_{\mu,\sigma^2}$ agrees with the full Gaussian compression model, each $\mathcal{N}_{\mu,1}$ agrees with the Gaussian compression model with variance 1, and each $\mathcal{N}_{0,\sigma^2}$ agrees with the Gaussian compression model with mean 0. The density (6) and the uniform density in the other two cases can be obtained from this agreement.

The example of the Gaussian model with variance 1 will be most useful for us in this paper (see Sect. 5 below). In the case of pivotal models it is clearly the simplest one among those that we discussed. In the case of online compression modelling, the summarizing statistic (5) is also the simplest one, but, unusually, the conditional distributions $P_n(\sigma_n)$ are not uniform (another such example is discussed in Vovk et al. 2022, Sect. 11.3.7).

Conformal testing produces unusual conformal test martingales for two reasons:

- these conformal test martingales are in a filtration that is poorer than the natural filtration generated by the observations $Z_1, Z_2, \ldots$ (we are forgetting some information);
- the martingales are *randomized*, in the sense of depending on the random values τ_n.

In this paper I will concentrate on the first reason (which appears to be more important). To get rid of the second reason, we will consider an online compression model that does not require it (in the sense that the p-values do not depend on the τs apart from the first one, p_1).

Remark 3.6. The expression "unnatural feature" used in Sect. 1 referred to the underlying filtration $(\mathcal{F}'_n)$ being different from the natural filtration.

4 General scheme of online testing

In this section we sketch (somewhat informally) a general testing scheme covering conformal testing and the other two approaches considered in the previous section. When processing the random observations $Z_1, Z_2, \ldots$ while testing $(P_\theta \mid \theta \in \Theta)$ as our null hypothesis, we proceed as follows.

1. We use a random number generator producing independent $\tau_1, \tau_2, \ldots$ that are uniformly distributed on $[0, 1]$; the sequence of τ is required to be independent of $Z_1, Z_2, \ldots$ for each $\theta \in \Theta$.
2. We then transform the sequence of observations $z_1, z_2, \ldots$ and each parameter value $\theta \in \Theta$ to $z_1^\theta, z_2^\theta, \ldots$: each z_n^θ is a function of θ and $z_1, \ldots, z_n$ that is measurable for each fixed value of θ (no measurability in θ is required). Typically this step reduces the information contained in $z_1, z_2, \ldots$ (for each θ).
3. Next, for each $\theta \in \Theta$, we gamble against the reduced sequence $z_1^\theta, z_2^\theta, \ldots$ and $\tau_1, \tau_2, \ldots$ obtaining a test martingale S^θ (in the reduced filtration extended by the τ) with respect to P_θ. Equivalently, we gamble against the extended observations (z_n^θ, τ_n). Our capital S_n^θ at time n is a function of $(z_1^\theta, \tau_1), \ldots, (z_n^\theta, \tau_n)$.
4. Finally, we use (1) as the amount of evidence found against (P_θ) at time n.

A natural question is whether this scheme is really general, but it does cover the three methods described in the previous section. These are special cases:

- In the simple element-wise testing scheme of Ramdas et al. [2022] (see Sect. 3.1), item 1 is not needed, the transformation in item 2 is identical (i.e., there is no transformation), and the gambling method in item 3 is to use likelihood ratios with P_θ in the denominator and the same probability measure (a mixture of P_θs) in the numerator.

- In neo-fiducial testing of Sect 3.2, item 1 is not needed. The transformation in item 2 and gambling in item 3 do not depend on θ, and item 4 is not needed.
- In conformal testing, the transformation in item 2 and gambling in item 3 do not depend on θ. Therefore, item 4 is not needed. For some online compression models, such as the Gaussian models discussed earlier, item 1 is also not needed (apart from the first few p-values).

5 Need for forgetting

Conformal testing often works well for testing the exchangeability model [Vovk et al., 2022, Part III]. On the other hand, it is obvious that no successful gambling is possible without forgetting against the null hypothesis of exchangeability: if under the null hypothesis there are no restrictions on the probability distribution of one observation, our capital can only go down (or stay at the same level). This is discussed in detail in Vovk et al. [2022, Sect. 8.6.1] and stated in Ramdas et al. [2022] as Theorem 17.

Therefore, it is essential to allow the test martingales in the element-wise scheme to depend on the value of the parameter θ if we want to avoid forgetting. In this section we will give an example where even such dependence does not solve the problem.

As before, we observe a sequence $Z_1, Z_2, \dots \in \mathbf{Z}$ generated by a probability measure in a family $(P_\theta \mid \theta \in \Theta)$, and we would like to have an online measure of evidence found against $(P_\theta \mid \theta \in \Theta)$ as null hypothesis. For each θ, we take a test martingale S^θ with respect to P_θ and the natural filtration $\mathcal{F} = (\mathcal{F}_n)$ (i.e., $\mathcal{F}_n$ is generated by $Z_1, \dots, Z_n$), and consider the element-wise test (1) as the amount of evidence found against $(P_\theta \mid \theta \in \Theta)$ at time n.

5.1 An example for pivotal testing

The following simple example shows the inadequacy of element-wise tests. We are testing the Gaussian pivotal model with variance 1, or the statistical model $(\mathcal{N}_{\mu,1} \mid \mu \in \mathbb{R})$. The normalizing transformation (3) acts on the random observations as

$$Z_1, Z_2, \dots \mapsto Z'_1, Z'_2, \dots,$$

where $Z'_n := Z_n - Z_1$, so that $Z'_n \sim \mathcal{N}_{0,2}$ for $n \geq 2$. Consider the process

$$S_n := \begin{cases} 1 & \text{if } n \leq 1 \\ 1/\mathcal{N}_{0,2}([-1,1]) & \text{if } n \geq 2 \text{ and } Z'_2 \in [-1,1] \\ 0 & \text{if } n \geq 2 \text{ and } Z'_2 \notin [-1,1]. \end{cases}$$

It can be considered both as a function of $Z_1, Z_2, \dots$ and as a function of $Z'_1, Z'_2, \dots$, but it is a martingale only as a function of $Z'_1, Z'_2, \dots$ (i.e., in the

reduced filtration $(\mathcal{F}'_n)$, where $\mathcal{F}'_n$ is generated by $Z'_1, \dots, Z'_n$). If we express it as a function of $Z_1, Z_2, \dots$, it becomes

$$S_n = \begin{cases} 1 & \text{if } n \le 1 \\ 1/\mathcal{N}_{0,2}([-1,1]) & \text{if } n \ge 2 \text{ and } Z_2 - Z_1 \in [-1,1] \\ 0 & \text{if } n \ge 2 \text{ and } Z_2 - Z_1 \notin [-1,1]. \end{cases} \tag{7}$$

Let us check that S is not covered by element-wise testing, i.e., $S_n \le \inf_\mu S_n^\mu$ does not hold for all n, S^μ being a natural test martingale (i.e., a test martingale for the natural filtration) with respect to $\mathcal{N}_{\mu,1}$. In fact, we will see that S_n is not dominated by any natural test martingale S_n^μ at times $n = 1$ and $n = 2$. Indeed, if it is, we must have

$$S_1^\mu(z_1) \ge \max\left(\frac{\mathcal{N}_{\mu,1}([z_1 - 1, z_1 + 1])}{\mathcal{N}_{0,2}([-1,1])}, 1\right)$$

for any $z_1 \in \mathbb{R}$. Notice that

$$\frac{\mathcal{N}_{\mu,1}([z_1 - 1, z_1 + 1])}{\mathcal{N}_{0,2}([-1,1])} > 1$$

holds for a nontrivial range of z_1: for example,

$$\frac{\mathcal{N}_{\mu,1}([\mu - 1, \mu + 1])}{\mathcal{N}_{0,2}([-1,1])} = \frac{\mathcal{N}_{0,1}([-1,1])}{\mathcal{N}_{0,2}([-1,1])} \approx 1.31 > 1.$$

Therefore, the expectation of $S_1^\mu(z_1)$ under $z_1 \sim \mathcal{N}_{\mu,1}$ must exceed 1, which contradicts S^μ being a test martingale.

5.2 An example for conformal testing

Our example for conformal testing will be a simple modification of the example for pivotal models given in the previous subsection. We again consider the Gaussian model with variance 1, but now its the online compression model with summarizing statistic (5) and conditional density (6). The equality

$$z_1^2 + z_2^2 = \frac{(z_1 + z_2)^2}{2} + \frac{(z_1 - z_2)^2}{2}$$

shows that $P_2(\sigma_2)$ generates (Z_1, Z_2) with $Z_2 - Z_1 \sim N(0, 2)$, and so

$$p_2 = \Phi((Z_2 - Z_1)/\sqrt{2})$$

(Φ standing for the standard Gaussian cumulative distribution function) if we choose $A(\sigma, z) := z$ as conformity measure.

Now we have the conformal test martingale

$$S_n := \begin{cases} 1 & \text{if } n \le 1 \\ 1/\mathcal{N}_{0,2}([-1,1]) & \text{if } n \ge 2 \text{ and } p_2 \in [\Phi(-1/\sqrt{2}), \Phi(1/\sqrt{2})] \\ 0 & \text{if } n \ge 2 \text{ and } p_2 \notin [\Phi(-1/\sqrt{2}), \Phi(1/\sqrt{2})] \end{cases}$$

in analogy with (7). As before, it is not dominated by any natural martingale S^μ for any $\mathcal{N}_{\mu,1}$.

6 Element-wise testing partially works for a fixed horizon

In this section we give theoretical results showing that the power of forgetting is limited, unfortunately in a very weak sense.

6.1 Finite horizon

We start from a simple result for a finite *horizon* N (i.e., we have only N observations, or are only interested in the first N observations).

Proposition 6.1. *Let $N \in \{1, 2, \dots\}$, and let (S^θ) be a family of test martingales, as in the scheme of Sect. 4 (perhaps not natural). Then there exists a family of natural test martingales $(\tilde{S}^\theta)$ such that*

$$\inf_{\theta\in\Theta} \tilde{S}^\theta_N = \inf_{\theta\in\Theta} S^\theta_N.$$

The natural element-wise test $\inf_\theta \tilde{S}^\theta_n$ in Proposition 6.1 can sometimes be less than the original element-wise test $\inf_\theta S^\theta_n$ at some time $n < N$, but it will eventually catch up (always, not just almost surely).

Proof. [Proof of Proposition 6.1] Let us fix a family of test martingales (S^θ). The expectation of our capital S^θ_N at step N is 1, and for each θ we get a natural test martingale $\tilde{S}^\theta_n$, $n \in \{0, 1, \dots, N\}$, by setting $\tilde{S}^\theta_N := S^\theta_N$ and averaging backwards:

$$\tilde{S}^\theta_n := \mathbb{E}^\theta(\tilde{S}^\theta_{n+1} \mid \mathcal{F}_n), \quad n = N-1, \dots, 0, \tag{8}$$

where $\mathbb{E}^\theta$ stands for the expectation with respect to P_θ. □

Suppose we are given a test martingale S that is not natural, such as the ones used in our counter-examples in Sect. 5.1 and Sect. 5.2. The problem is that for steps before N the backward averaging (8) may give a result different from (and therefore not dominating) S_n, $n < N$.

A disadvantage of Proposition 6.1 is that it ignores the complexity, in any sense (computational, descriptional, etc.), of the natural test martingale $\tilde{S}$. While S may be very easy to define and independent of θ, such as a Composite Jumper conformal test martingale [Vovk et al., 2022], $\tilde{S}$ will depend on θ and may be much more complicated.

The $\inf_\theta$ in the definition (1) of element-wise testing does not even have to be measurable, as we already mentioned, and in some sense it is not even well-defined (for each of the uncountably many θ the corresponding test martingale is defined to within a P_θ-null set, which makes the $\inf_\theta$ undefined under standard definitions. (For rich spaces Θ and Ω, we can even make $\inf_\theta \tilde{S}^\theta_n = 0$, $n < N$, by an awkward choice of a version of the conditional expectation in (8).)

The idea in the proof of Proposition 6.1 can also be applied to randomized test martingales (such as conformal test martingales under the exchangeability

model). Suppose $S_n = S_n^\theta$ does not depend on θ (as for conformal test martingales). We can then average S_N with respect to the random numbers $\tau_1, \dots, \tau_N$ and after that apply backward averaging with respect to the σ-algebras $\mathcal{F}_n$:

$$\tilde{S}_n^\theta := \mathbb{E}^\theta \left(\mathbb{E}^\tau (S_N) \mid \mathcal{F}_n \right), \tag{9}$$

where, of course, $\mathbb{E}^\tau$ refers to averaging over the random numbers (produced independently from the uniform distribution on $[0, 1]$). We will then have

$$\inf_{\theta \in \Theta} \tilde{S}_N^\theta = \mathbb{E}^\tau (S_N).$$

6.2 Infinite horizon

Another disadvantage of Proposition 6.1 is that it is only applicable to a finite horizon. We can generalize it by allowing N to be, e.g., a bounded stopping time, but a natural question is whether it holds asymptotically at infinity for the infinite horizon $N := \infty$. The next proposition is a step in this direction, but it is very restrictive (as we will discuss momentarily).

Proposition 6.2. *Suppose that the parameter set Θ is finite and that different P_θ in the statistical model (our null hypothesis) (P_θ) are mutually singular. Let (S^θ) be a family of test martingales, and let $\epsilon > 0$ (be arbitrarily small). Then there exists a family of natural test martingales $(\tilde{S}^\theta)$ such that*

$$\liminf_{n \to \infty} \inf_\theta \tilde{S}_n^\theta \geq (1 - \epsilon) \limsup_{n \to \infty} \inf_\theta S_n^\theta \tag{10}$$

a.s under any probability measure P_θ from the null hypothesis.

As in the case of Proposition 6.1, Proposition 6.2 says that, even when a natural test martingale $\tilde{S}^\theta$ falls below the original test martingale S^θ, it will eventually catch up (or at least almost catch up, to within any ϵ on the relative scale). The most restrictive condition in Proposition 6.2 is that Θ is finite (although it can be as dense as we wish).

Another restrictive condition in Proposition 6.2 is that different P_θ are required to be mutually singular. This condition often holds for interesting statistical models; for example, in the IID case it follows from Kakutani's theorem [Kakutani, 1948] that P_θ corresponding to different θ are either identical or mutually singular. Moreover, we can often even identify θ in the limit almost surely given a sequence observations generated from P_θ (formally, there exists a strongly consistent estimator for θ).

In (10) we have lim inf and lim sup instead of just lim. For the lim inf it is not essential, and we can replace it by lim, meaning that the limit will exist almost surely (although it can be ∞). Having lim sup is essential, but let me discuss it at the end of the proof.

Proof. [Proof of Proposition 6.2] By Doob's convergence theorem [Shiryaev, 2019, Corollary 7.4.3],

$$S_\infty^\theta := \lim_{n \to \infty} S_n^\theta \tag{11}$$

exists almost surely under P_θ. Without loss of generality we assume that its $\mathbb{E}^\theta$-expectation is 1 (its expectation is at most 1 by Fatou's lemma, and we can scale it up if the expectation is below 1). For each $\theta \in \Theta$ let us define the natural test martingale

$$\tilde{S}_n^\theta := \mathbb{E}^\theta(S_\infty^\theta \mid \mathcal{F}_n) \tag{12}$$

by backward induction; remember that this process is a test martingale only under P_θ. By Lévy's theorem [Shiryaev, 2019, Theorem 7.4.3] we have

$$\tilde{S}_n^\theta \to S_\infty^\theta$$

a.s. under P_θ. Since this convergence holds only P_θ-almost surely, we need to "regularize" $\tilde{S}^\theta$ to ensure its desired behaviour under $P_{\theta'} \neq P_\theta$.

For each pair $\theta, \theta' \in \Theta$ with $P_\theta \neq P_{\theta'}$, fix a natural test martingale $S^{\theta,\theta'}$ with respect to P_θ such that

$$\liminf_{n\to\infty} S^{\theta,\theta'} = \infty \quad P_{\theta'}\text{-a.s.}$$

Such a test martingale can be defined as the likelihood ratio $\mathrm{d}P_{\theta'} / \mathrm{d}P_\theta$ if $P_{\theta'}$ is locally absolutely continuous with respect to P_θ (see Shiryaev 2019, Theorem 7.6.2) and as an obvious modification of the likelihood ratio otherwise. Now we can redefine

$$\tilde{S}_n^\theta := (1-\epsilon)\tilde{S}_n^\theta + \frac{\epsilon}{|\Theta|-1} \sum_{\theta' \in \Theta \setminus \{\theta\}} S_n^{\theta,\theta'} \tag{13}$$

(assuming, without loss of generality, that $|\Theta| > 1$).

Under P_θ, we have, a.s.,

$$\lim_{n\to\infty} \tilde{S}_n^\theta \geq (1-\epsilon) S_\infty^\theta = (1-\epsilon) \lim_{n\to\infty} S_n^\theta, \tag{14}$$

and under $P_{\theta'}$ such that $P_{\theta'} \neq P_\theta$, we have, a.s.,

$$\liminf_{n\to\infty} \tilde{S}_n^\theta \geq \frac{\epsilon}{|\Theta|-1} \liminf_{n\to\infty} S_n^{\theta,\theta'} = \infty. \tag{15}$$

Combining (14) and (15), we obtain, almost surely under any element of the statistical model,

$$\liminf_{n\to\infty} \min_{\theta\in\Theta} \tilde{S}_n^\theta = \min_{\theta\in\Theta} \liminf_{n\to\infty} \tilde{S}_n^\theta \geq (1-\epsilon) \min_{\theta\in\Theta} \limsup_{n\to\infty} S_n^\theta.$$

Now we can discuss in detail the role of the lim inf and lim sup in (10). The lim inf can be replaced by lim since the latter exists (almost surely) under the true P_θ by Doob's convergence theorem and is ∞ under all other P_θ because of the components $S_n^{\theta,\theta'}$ in (13). As for the lim sup, $\lim_{n\to\infty} S_n^\theta$ exists almost surely under the true P_θ, but there are no constraints on S_n^θ's behaviour under the other P_θ; therefore, it is essential to have lim sup (unless we are willing to weaken (10) by replacing the lim sup with lim inf, or vice versa). □

6.3 Creating natural test martingales out of likelihood ratios

So far in this section we were discussing creating natural test martingales out of other test martingales (those over a reduced filtration). But the process has a bottleneck: first we define an e-variable (i.e., a nonnegative random variable with expectation 1, such as (11)) and then average it with respect to a filtration of σ-algebras (as in (12)). An easier option is to start directly from an e-variable over the first N observations, in the case of a finite horizon N.

When testing an online compression model (such as exchangeability), it serves as our null hypothesis. We also fix an *alternative hypothesis*, which, in the simplest case, is a probability measure Q on the sample space. (It can be a mixture of a Bayesian model, as in Vovk et al. 2022, Sect. 9.2.)

What is really important for us is not Q itself, but a regular conditional probability generated from Q, which in fact carries less information than Q does. (See, e.g., Rogers and Williams 2000, Sect. II.89, for a standard theorem about the existence of a suitable regular conditional probability.) Informally, we let q to be a Markov kernel mapping each $\sigma \in \Sigma(\mathbf{Z}^N)$ to a probability measure $q(\sigma)$ on the set $t_N^{-1}(\sigma)$.

To obtain a martingale from an alternative hypothesis, in the case of a finite horizon N, we can proceed as in Proposition 6.1, namely we set

$$S_n^\theta := \mathbb{E}^\theta \left(\frac{\mathrm{d}q(\sigma_N)}{\mathrm{d}P_N(\sigma_N)} \mid \mathcal{F}_n \right), \quad \theta \in \Theta, \quad n = 0, \dots, N. \tag{16}$$

We will see some experimental results for the final value

$$\frac{\mathrm{d}q(\sigma_N)}{\mathrm{d}P_N(\sigma_N)} \tag{17}$$

(which we call the *batch benchmark*) of this test martingale in the next section.

In the case of an infinite horizon and in the spirit of Proposition 6.2, we have an open problem. Consider the sequence of the summaries $\sigma_1, \sigma_2, \dots$ generated from the alternative hypothesis. For each of them define the likelihood ratio martingale (16). Do these test martingales converge? If so, can the limit be used for hypothesis testing?

7 Illustration: the problem of change detection

In this section I will illustrate several points raised in the previous sections using a simple example of changepoint detection with a finite horizon. Chapter 9 of Vovk et al. [2022] shows numerous examples where conformal testing (usually implemented as the Bayes–Kelly, or *BK*, conformal test martingale) is very close to natural benchmarks. In the example considered in this section the difference is deliberately made more pronounced. Namely, the observation space is $\mathbf{Z} := \{0, 1\}$, the null hypothesis is the randomness model $(\mathbf{B}_\theta^{20} \mid \theta \in [0, 1])$, $\mathbf{B}_\theta$ being the Bernoulli distribution on $\{0, 1\}$ with parameter θ, the alternative

hypothesis Q is that $N_0 := 10$ observations are generated from the Bernoulli distribution on $\{0, 1\}$ with parameter (probability of 1) $\pi_0 := 0.1$, and another $N_1 := 10$ observations are generated from the Bernoulli distribution on $\{0, 1\}$ with parameter $\pi_1 := 0.9$. All 20 observations are generated independently. In this setting the time horizon is finite and very short, 20.

We will only be interested in the final values of our martingales and related processes; for some of these processes the intermediate values are easily computable, but for others this is tricky (and requires further research). The final values are shown in Fig. 1, which will be explained in the rest of this section.

The boxplots in Fig. 1 represent results of 10^6 simulations of the final values (at time horizon 20) of five processes, including the Bayes–Kelly martingale (BK). Each boxplot shows the median as the horizontal orange line in the middle of the box, with notches representing a confidence interval around the median, the mean as a green triangle, the interquartile range as a box, and the 5% and 95% quantiles as whiskers.

The second process, "mean BK", is an approximation to the expectation (9) of the BK test martingale S_n. To compute its final value, we compute the final values of 10^3 simulations of the BK test martingale and then average them. If we average S_n for each $n = 0, \dots, N$, the resulting process will not be a martingale, as discussed in Vovk et al. [2022, Sect. 9.3]. However, the final value will be close to the final value of the expected BK test martingale (9) (and so is a valid measure of evidence collected against the null hypothesis).

The last two processes in Fig. 1 are the *lower benchmark* (LB) and the *upper benchmark* (UB). The former is

$$\mathrm{LB}_n := \inf_\theta \frac{Q(Z_1 = z_1, \dots, Z_n = z_n)}{\mathbf{B}_\theta(\{z_1\}) \dots \mathbf{B}_\theta(\{z_n\})},$$

where $z_1, \dots, z_n$ are the realized values of the random observations $Z_1, \dots, Z_n$,

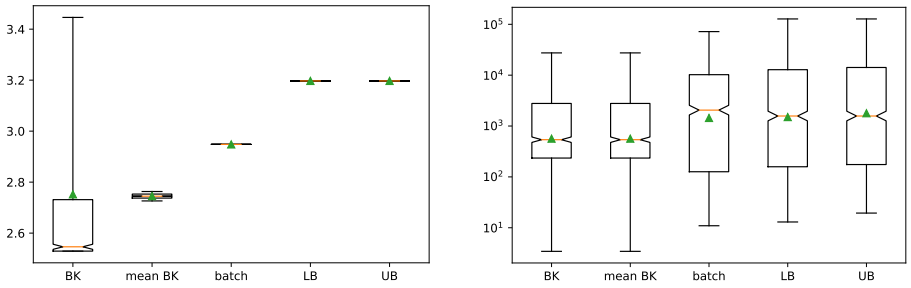

Figure 1: Left panel: Five final values as described in text for a fixed dataset (for a changepoint detection problem). Right panel: Five final values as described in text for random datasets.

respectively, and the latter is

$$\mathrm{UB}_n := \frac{Q(Z_1 = z_1, \dots, Z_n = z_n)}{\mathbf{B}_{0.5}(\{z_1\}) \dots \mathbf{B}_{0.5}(\{z_n\})}.$$

The lower benchmark is not a martingale under any $\mathbf{B}_\theta^{20}$, but it is a valid measure of evidence against the null since for each θ it is dominated by the likelihood ratio $Q(Z_1 = z_1, \dots, Z_n = z_n)/(\mathbf{B}_\theta(\{z_1\}) \dots \mathbf{B}_\theta(\{z_n\}))$. On the other hand, the upper benchmark is only valid under $\mathbf{B}_{0.5}^{20}$ (which is, in a sense, the mid-point between the pre-change and post-change distributions), and not valid under the other elements of the null hypothesis.

The middle process, the one labelled "batch", is a new benchmark (which we called the "batch benchmark" earlier), and we will discuss it at the end of this section.

The left panel of Fig. 1 shows the final value of the BK martingale, mean BK martingale, batch benchmark, and upper and lower benchmarks for a specific randomly generated dataset (using our default seed 42 for the random number generator). With a large probability, the number of 1s in the dataset will be 10, and so the upper and lower benchmarks will in fact coincide, as they do in Fig. 1.

The final value of the averaged BK martingale is higher than that of the BK martingale in the left panel of Fig. 1, and it is less volatile. It is higher because averaging on the log scale is akin to taking maximum, as we pointed out in Vovk et al. [2022, Sect. 9.3]. It is clear that the genuine average (expectation) of the BK martingale over the random numbers τ is even higher (with zero volatility), but it is only marginally higher (as our other experiments show).

If the dataset is randomized, the difference is much less noticeable: see the right panel of Fig. 1. In particular, the difference between BK and mean BK is swamped by the variability due to the random choice of a dataset. The three benchmarks, however, are still significantly higher.

Now let us spell out the batch benchmark (17), shown in the middle boxplots of both panels of Fig. 1, for this case. Suppose the observed data sequence is $z_1, \dots, z_N$ and let

$$K := \sum_{n=1}^{N} z_n, \quad k_0 := \sum_{n=1}^{N_0} z_n, \quad k_1 := \sum_{n=N_0+1}^{N_1} z_n$$

be the number of 1s among all observations, among the pre-change observations, and among the post-change observations, respectively. The probability of $z_1, \dots, z_N$ under the alternative is

$$\pi_0^{k_0}(1-\pi_0)^{N_0-k_0}\pi_1^{k_1}(1-\pi_1)^{N_1-k_1}$$

and the number of data sequences leading to the same k_0 and k_1 is

$$\binom{N_0}{k_0}\binom{N_1}{k_1}.$$

The exchangeability summary (i.e., the summary under the exchangeability compression model) of $z_1, \dots, z_N$ is K, and so the conditional probability of $z_1, \dots, z_N$ given its exchangeability summary under the alternative hypothesis is

$$\frac{\pi_0^{k_0}(1-\pi_0)^{N_0-k_0}\pi_1^{k_1}(1-\pi_1)^{N_1-k_1}}{\sum_{k=(K-N_0)^+}^{K\wedge N_1}\binom{N_0}{K-k}\binom{N_1}{k}\pi_0^{K-k}(1-\pi_0)^{N_0-K+k}\pi_1^{k}(1-\pi_1)^{N_1-k}} = \frac{1}{\sum_{k=(K-N_0)^+}^{K\wedge N_1}\binom{N_0}{K-k}\binom{N_1}{k}\left(\frac{(1-\pi_0)\pi_1}{\pi_0(1-\pi_1)}\right)^{k-k_1}}.$$

In this formula, k is the analogue of k_1 for the other elements of $t_N^{-1}(\sigma_N)$ (where t_N and σ_N refer to the exchangeability model), and $K-k$ is the analogue of k_0. It is clear that k ranges from $(K-N_0)^+$ (where $u^+ := \max(u,0)$) and $K \wedge N_1$ (where $u \wedge v := \min(u,v)$); it is easy to check directly that $(K-N_0)^+ \le K \wedge N_1$. The conditional probability of $z_1, \dots, z_N$ given its exchangeability summary under the null hypothesis is

$$1/\binom{N}{K},$$

which gives the explicit expression

$$\frac{\binom{N}{K}}{\sum_{k=(K-N_0)^+}^{K\wedge N_1}\binom{N_0}{K-k}\binom{N_1}{k}\left(\frac{(1-\pi_0)\pi_1}{\pi_0(1-\pi_1)}\right)^{k-k_1}}$$

for the batch benchmark (17) that we use in our experiments.

The right panel of Fig. 1 shows that the batch benchmark is competitive with the lower and upper benchmarks. It looks a promising option. Its advantage over the upper benchmark is obvious: it is valid under any power probability measure, not just under $\mathbf{B}_{0.5}^N$. One advantage over the lower benchmark is that it is admissible for each parameter value θ, whereas the inadmissibility of the lower benchmark for some θ is obvious.

8 Conclusion

I have mentioned several directions of further research in the previous sections, but these are a few more:

- In Sect. 5.2 we saw that for the model $(\mathcal{N}_{\mu,1})$ the element-wise tests are inadequate. It would be interesting to quantify this observation and to extend it to other online compression models.
- Can we apply explicitly Proposition 6.1 and (9) to get explicit expressions for the natural modifications of the numerous conformal test martingales described in Vovk et al. [2022, Part III]?

- Relaxing the assumptions of Proposition 6.2 (such as Θ being finite) or showing that it is impossible.
- In Sect. 7 we only studied the final values of various test martingales. Their intermediate values deserve to be studied both theoretically and experimentally.

A characteristic feature of conformal testing is that part of the data is forgotten in the process of gambling against the null hypothesis (such as exchangeability). On the other hand, the same test martingale works against every probability measure in the null hypothesis. We have seen that forgetting is essential, even if our gambling strategy is allowed to depend on a probability measure in the null hypothesis.

We saw that we can get rid of forgetting, but to a very limited extent. It is not clear at all how the power and versatility of conformal testing can be achieved without forgetting, and it appears that, at least for the time being, we should embrace forgetting and live with it.

Acknowledgement

Thanks to Wouter Koolen for an illuminating discussion of element-wise testing as formalization of online testing during ISIPTA 2019 (3–6 July 2019). Thanks to Aaditya Ramdas for another illuminating discussion during a meeting in Amsterdam between the authors of Ramdas et al. [2023] hosted by Peter Grünwald (4–6 June 2022). While the first discussion concentrated on advantages of element-wise testing, the second one was all about their limitations.

References

Prasanta S. Bandyopadhyay and Malcolm R. Forster, editors. *Philosophy of Statistics*. Elsevier, Amsterdam, 2011.

George A. Barnard. Sequential tests in industrial statistics (with discussion). *Supplement to the Journal of the Royal Statistical Society*, 8:1–26, 1946.

George A. Barnard. "Sequential Analysis" by Abraham Wald: Review. *Journal of the American Statistical Association*, 42:658–665, 1947.

Laurent Bienvenu, Peter Gács, Mathieu Hoyrup, Cristobal Rojas, and Alexander Shen. Algorithmic tests and randomness with respect to a class of measures. *Proceedings of the Steklov Institute of Mathematics*, 274:34–89, 2011.

Mike Byster. *The Power of Forgetting*. Harmony, New York, 2014.

David R. Cox. Some problems connected with statistical inference. *Annals of Mathematical Statistics*, 29:357–372, 1958.

Ronald A. Fisher. Applications of "Student's" distribution. *Metron*, 5:90–104, 1925a.

Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1925b.

Ronald A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

Peter Gács. Lecture notes on descriptional complexity and randomness. Technical Report arXiv:2105.04704 [cs.IT], arXiv.org e-Print archive, May 2021.

Ian Hacking. *Logic of Statistical Inference*. Cambridge University Press, Cambridge, 1965.

Shizuo Kakutani. On equivalence of infinite product measures. *Annals of Mathematics*, 49:214–224, 1948.

Peter McCullagh, Vladimir Vovk, Ilia Nouretdinov, Dmitry Devetyarov, and Alex Gammerman. Conditional prediction intervals for linear regression. In *Proceedings of the Eighth International Conference on Machine Learning and Applications (ICMLA 2009)*, pages 131–138, 2009. Available from http://www.stat.uchicago.edu/~pmcc/reports/predict.pdf.

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.

Karl Pearson. On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175, 1900.

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.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 2023. Under revision; a preliminary version published as arXiv:2210.01948 [math.ST].

L. Chris G. Rogers and David Williams. *Diffusions, Markov Processes, and Martingales*. Cambridge University Press, Cambridge, second edition, 2000.

Tore Schweder and Nils L. Hjort. *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge University Press, Cambridge, 2016.

Glenn Shafer and Vladimir Vovk. *Probability and Finance: It's Only a Game!* Wiley, New York, 2001.

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

Jieli Shen, Regina Liu, and Minge Xie. Prediction with confidence—a general framework for predictive inference. *Journal of Statistical Planning and Inference*, 195:126–140, 2018.

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

Student (William S. Gosset). The probable error of a mean. *Biometrika*, 6: 1–25, 1908.

Jean Ville. *Etude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.

Vladimir Vovk. On the concept of the Bernoulli property. *Russian Mathematical Surveys*, 41:247–248, 1986.

Vladimir Vovk and Vladimir V. V'yugin. On the empirical validity of the Bayesian method. *Journal of the Royal Statistical Society B*, 55:253–266, 1993.

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.

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

Abraham Wald. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, 16:117–186, 1945.

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

Minge Xie and Kesar Singh. Confidence distribution, the frequentist distribution estimator of a parameter: a review. *International Statistical Review*, 81: 3–39, 2013.