Optimal test statistics for anytime-valid hypothesis tests

Lardy, T.D.

7 | Tests of Group Invariance

One of the key assumptions underlying our approach to the construction of anytime-valid tests in the previous chapters was that the data were independent. We now consider a setting where this assumption does not necessarily hold. In particular, we study the problem of testing for invariance under a group of transformations, which includes many standard statistical hypothesis tests, such as those for normality and exchangeability. We show that, regardless of any dependence structure in the original data, the invariance properties of the problem can be used to construct a sequence of random variables that are i.i.d. under the null hypothesis. In fact, these transformed data always have the same distribution under the null hypothesis, that is, the null becomes simple. Consequently, it is straightforward to compute the log-optimal e-statistic for a given alternative, as it is simply given by the likelihood ratio between the alternative and (now simple) null. It can be shown that the cumulative product of these log-optimal e-statistics, the likelihood ratio process, is the log-optimal e-process among all e-processes that are functions of the transformed data. Remarkably, under some assumptions on the alternative, it is sometimes even log-optimal among all e-processes (that is, which are functions of the original data).

We furthermore apply this method to extend recent anytime-valid tests of independence, which leverage exchangeability, to work under general group invariances. Additionally, we show applications to testing for invariance under subgroups of rotations, which corresponds to testing the Gaussian-error assumptions behind linear models.

7.1 Introduction

Symmetry plays a crucial role in statistical modeling. Most models, either explicitly or implicitly, introduce assumptions of distributional symmetry about the data. For example, any distribution under which data are independent and identically distributed is symmetric under permutations of the data points and any regression model with Gaussian errors is symmetric under certain rotations of the data. If these symmetries are actually present in the data, employing symmetric models yields advantages for various objectives. These objectives include max-min optimality in hypothesis tests (Lehmann and Romano, 2005) (see also Chapter 8), admissibility of estimators (Brown, 1966), and increased predictive performance of neural networks (Cohen and Welling, 2016). On the other hand, if the symmetries are absent from the data, the use of a symmetric model may lead to poor performance in these same tasks. We address the problem of testing for the presence of symmetries in the data.

The presence of a symmetry is formalized as a null hypothesis of distributional invariance under the action of a group (in the algebraic sense). Perhaps the most prominent example is infinite exchangeability—the hypothesis that the distribution of any finite data sequence is invariant under the group of all permutations. The null hypothesis of exchangeability is at the heart of classic methods such as permutation tests (Fisher, 1936; Pitman, 1937) and rank tests (Sidak et al., 1999). Tests for other symmetries have also been studied, including tests for rotational symmetry, which corresponds to invariance under the orthogonal group (Baringhaus, 1991), symmetries for data taking values on groups (Diaconis, 1988), and more general frameworks (Lehmann and Stein, 1949; Chiu and Bloem-Reddy, 2023). The majority of tests in this line of work are designed for fixed-sample experiments—the amount of data to be collected is determined before the experiment. In this chapter, we focus on testing for the presence of symmetries sequentially and under continuous monitoring.

In the applications that interest us, data are analyzed as they are collected, and the decisions to either stop and reach a conclusion or to continue data collection may depend on what has been observed so far. Hypothesis tests that retain type-I error control under such flexible data collection schemes have been called tests of power one (Robbins and Siegmund, 1974; Lai, 1977), and, more recently, anytime-valid tests (Ramdas et al., 2023). The main insights in this line of work are that a test martingale—a nonnegative martingale with expected value equal to one—can be monitored continuously, and that a test that rejects when the test martingale exceeds a fixed threshold maintains type-I error control uniformly over time (Ramdas et al., 2020; Shafer, 2021). More generally, the minimum over a family of test martingales—an e-process—can be monitored (Ramdas et al., 2020).

While anytime-valid tests of general symmetries have not received much attention, the specific case of infinite exchangeability has been studied classically. For example, sequential rank tests, which can be interpreted as tests of exchangeability (also called tests of randomness in the literature), have been studied. Sen and Ghosh (1973a,b, 1974) develop asymptotic approximations and law-of-the-iterated-logarithm-type inequalities for linear rank statistics that hold uniformly over the duration of the experiment. More recently, Ramdas et al. (2022) and Saha and Ramdas (2024) developed e-processes for the hypothesis of infinite exchangeability under specific assumptions (binary and paired data, respectively). Anytime-valid tests of exchangeability that do not require any additional assumptions are addressed in the work on conformal prediction (Vovk et al., 2003, 2005). Conformal prediction, perhaps best known as a framework for uncertainty quantification for point predictors, can also be used to produce test martingales to test for exchangeability. In the context of conformal prediction, test martingales are called conformal martingales. Most crucially for our present purposes, Vovk et al. (2005) show that conformal martingales cannot only be used to test for infinite exchangeability, but also to test whether data are generated by a fully general class of sequential data-generating mechanisms, called online compression models (see Section 7.3). It is natural to ask whether distributionally symmetric models define online compression models, as the conformal martingales built by Vovk et al. (2005) would automatically yield tests of distributional symmetry. Unfortunately, this is not true in general.

In this chapter, we show that the above difficulty can be circumvented: Under natural conditions, a distributionally symmetric model does define an online compression model. Furthermore, we show that the resulting conformal martingales are optimal in a specific sense. Indeed, we show that the resulting martingales are likelihood ratios against implicit alternatives and prove that they are optimal—in a sense that is specified in Section 7.2—for testing against that particular alternative. We use these constructions to abstract and generalize existing tests of independence under the assumption of exchangeability (Henzi and Law, 2024) to tests of independence under general symmetries. Finally, we build tests for the Gaussian-error assumptions behind linear models by testing for invariance under subgroups of the orthogonal group.

The rest of this document is organized as follows. Section 7.2 formally introduces the problem of anytime valid testing for distributional invariance, and the optimality criterion that we employ. Then, in Section 7.3, the connection between group-invariant distributions and online compression models is shown. This connection is used in Section 7.4 to construct test martingales against the hypothesis of distributional invariance; the optimality of this procedure is shown in Section 7.4.2. Section 7.5 shows applications to test the assumptions of linear models, testing sign-invariant exchangeability, and independence testing. Finally, Section 7.6 discusses a potential direction for future work.

7.2 Problem Statement

Suppose that we observe data $X_1, X_2, \ldots$ sequentially and that they take values in some topological space $\mathcal{X}$. In our examples, $\mathcal{X} = \mathbb{R}$. Note that we assume neither that these observations are independent nor that they are identically distributed; we only assume that they are sampled from a distribution on infinite sequences. Furthermore, for each $n = 1, 2, \ldots$, we assume that G_n is a compact topological group (in the algebraic sense) that acts continuously on $\mathcal{X}^n$. Here, a topological group is a group that is equipped with a topology under which the group operation, seen as a function $G_n \times G_n \to G_n$, is a continuous map. A (left) group action is a map $\varphi : G_n \times \mathcal{X}^n \to \mathcal{X}^n$ that satisfies, for any $g, h \in G_n$ and $x^n \in \mathcal{X}^n$, that $\varphi(h, \varphi(g, x^n)) = \varphi(hg, x^n)$. To alleviate notation, when the action is clear from context, we write gx^n instead of $\varphi(g, x^n)$. In our examples, the group G_n has a representation as a group of $n \times n$ matrices and the group acts on $\mathbb{R}^n$ by matrix multiplication. We are interested in testing the null hypothesis of invariance of the data under the action of the sequence of groups $(G_n)_{n \in \mathbb{N}}$, that is,

$$\mathcal{H}_0 : gX^n \stackrel{\mathcal{D}}{=} X^n \quad \text{for all } g \in G_n \text{ and all } n \in \mathbb{N}, \tag{7.1}$$

where $X^n = (X_1, \ldots, X_n)$ and $\stackrel{\mathcal{D}}{=}$ signifies equality in distribution. At this level of generality, one can build pathological examples of (7.1) that cannot be tested; more structure is needed (see Section 7.3). The next example contains simple instances of the problems that are amenable to our general framework.

Example 7.1 (Exchangeability, rotational symmetry, and compact matrix groups)**.** For tests of infinite exchangeability, the null hypothesis is given by

$$\mathcal{H}_0 : X_1, \ldots, X_n \text{ are exchangeable for each } n \in \mathbb{N}.$$

By definition, this can be rewritten as

$$\mathcal{H}_0 : (X_{\pi(1)}, \ldots, X_{\pi(n)}) \overset{\mathcal{D}}{=} (X_1, \ldots X_n) \text{ for all } \pi \in S(n) \text{ and } n \in \mathbb{N},$$

where $S(n)$ denotes the group of permutations on n elements. In terms of the notation above, the relevant sequence $(G_n)_{n\in\mathbb{N}}$ of groups is $G_n = S(n)$ and the group action may be written as $\pi X^n = (X_{\pi(1)}, \ldots, X_{\pi(n)})$ for each permutation $\pi \in S(n)$. Note that $S(n)$ can be represented through the group of $n \times n$ permutation matrices (matrices with exactly one entry of 1 in each row and each column and 0 in all other entries). The hypothesis above coincides with that of distributional invariance under multiplication of the data $(X_1, \ldots, X_n)$ by $n \times n$ permutation matrices for all n.

Similarly, in the case of tests for sphericity, that is, invariance under rotations of data, the relevant sequence of groups is $G_n = \mathrm{O}(n)$. Here, $O(n)$ denotes the orthogonal group—the group of all $n \times n$ matrices O with orthonormal columns, that is, such that $O^T O = I$. Details are given in Section 7.5.2. The action of permutation and orthogonal groups are special examples of the actions of classic compact matrix groups on $\mathbb{R}^n$ (Meckes, 2019). With adjustments, invariance under any of these classic compact matrix groups is also an instance of the hypothesis in 7.1.

Anytime-valid tests We are interested in constructing sequential tests for $\mathcal{H}_0$ as in (7.1) that are anytime-valid at some prescribed level $\alpha \in (0, 1)$. Here, a sequential test is a sequence $(\varphi_n)_{n\in\mathbb{N}}$ of rejection rules $\varphi_n : \mathcal{X}^n \to \{0, 1\}$ and we say that it is anytime valid for $\mathcal{H}_0$ at level α if

$$Q(\exists n \in \mathbb{N} : \varphi_n = 1) \le \alpha \text{ for any } Q \in \mathcal{H}_0.$$

Notice that this is a type-I error guarantee that is valid uniformly over all sample sizes: the probability that the null hypothesis is ever rejected by $(\varphi_n)_{n\in\mathbb{N}}$ is controlled by α. The main tools for constructing anytime-valid tests are test martingales (Ramdas et al., 2020; Shafer, 2021; Grünwald et al., 2024) and minima thereof, e-processes—see Ramdas et al. (2020) for a comprehensive overview. We now define them.

Test martingales A sequence of statistics of the data is a test martingale if it is non-negative, starts at one, and is a supermartingale under every element of $\mathcal{H}_0$. Formally, let $\mathbb{G} = (\mathcal{G}_n)_{n\in\mathbb{N}}$ be a filtration of σ-algebras such that $\mathcal{G}_n \subseteq \sigma(X^n)$, where $\sigma(X^n)$ denotes the σ-algebra induced by X^n. Then a sequence of statistics $(M_n)_{n\in\mathbb{N}}$ that is adapted to $\mathbb{G}$ is a test martingale for $\mathcal{H}_0$ with respect to $\mathbb{G}$ if $\mathbf{E}_Q [M_n \mid \mathcal{G}_{n-1}] \le M_{n-1}$

for all $Q \in \mathcal{H}_0$ and $M_0 = 1$. The main utility of test martingales is that, under $\mathcal{H}_0$, they take large values with small probability. This is quantified by Ville's inequality (Ville, 1939), which shows that the sequential test given by $\varphi_n = \mathbf{1}\{M_n \geq 1/\alpha\}$ is anytime valid.

Lemma 7.1 (Ville's inequality). *Let $(M_n)_{n\in\mathbb{N}}$ be a test martingale with respect to some filtration $(\mathcal{G}_n)_{n\in\mathbb{N}}$ under all elements of $\mathcal{H}_0$, then*

$$\sup_{Q\in\mathcal{H}_0} Q(\exists n \in \mathbb{N} : M_n \geq 1/\alpha) \leq \alpha.$$

Proof. Fix $Q \in \mathcal{H}_0$. Doob's optional stopping theorem states that $\mathbb{E}_Q[M_\tau] \leq 1$ for any stopping time τ that is adapted to $(\mathcal{G}_n)_{n\in\mathbb{N}}$ (Durrett, 2019, Theorem 5.7.6). Markov's inequality implies that $Q(M_\tau > \frac{1}{C}) \leq C$ for any $C > 0$. Applying this to the stopping time $\tau^* = \inf\{n \in \mathbb{N} : M_n \geq \frac{1}{\alpha}\}$ shows the result. □

Test martingales that make use of external randomization will also prove useful; we will call them randomized test martingales. For randomized test martingales, we append an independent random number $\theta_n \sim \text{Uniform}([0,1])$ to each X_n, that is, we let $Y_n = (X_n, \theta_n)$ and consider test martingales that are functions of Y_n rather than X_n.

Test martingales are part of a broader class of processes, e-processes (Ramdas et al., 2023). An e-process is any nonnegative stochastic process E such that $\mathbb{E}_Q[E_\tau] \leq 1$ for all $Q \in \mathcal{H}_0$ and (a subset of) all stopping times τ. Any e-process can be turned into an anytime-valid test by thresholding it, that is, $\phi_n = \mathbf{1}\left\{E_\tau \geq \frac{1}{\alpha}\right\}$ is an anytime-valid test for any stopping time τ. This property is often referred to as safety under optional stopping (Grünwald et al., 2024). Relatedly, the product of e-processes based on independent data is again an e-process. That is, suppose some e-processes E and E' are used for independent experiments, yielding stopped process E_{τ_1} and E'_{τ_2}. Then the product again has the property that $\mathbb{E}_Q[E_{\tau_1} E'_{\tau_2}] \leq 1$ for all $Q \in \mathcal{H}_0$, which is referred to as safety under optional continuation.

Log-optimality This type of evidence aggregation by multiplication of e-processes motivates a natural optimality criterion. Indeed, suppose we were to repeatedly run a single experiment, using a fixed e-process E and stopping time τ. If we measure the total evidence by the cumulative product of the individual e-processes, then the asymptotic growth rate of our evidence under true distribution P will be $\mathbb{E}_P[\log E_\tau]$. It is therefore custom to look for e-processes that maximize this asymptotic growth rate, as can be traced back to Kelly betting (Kelly, 1956). Variants of this criterion have more recently been studied under numerous monikers (Shafer, 2021; Koolen and Grünwald, 2022; Grünwald et al., 2024), but here we shall simply refer to maximizers of this criterion as "log-optimal".

7.3 Sequential Group Actions are Online Compression Models

The hypothesis in (7.1) can only be meaningfully tested if the statements regarding group invariance for each $n \in \mathbb{N}$ are consistent with each other; without any further restrictions, invariance of the data at one time may contradict the invariance of the data at a later time. To avoid such situations, we assume that there is a certain structure to the action of sequence of groups $(G_n)_{n\in\mathbb{N}}$ on the sample space, which we will refer to as a sequential group action. After the statement of this definition, we discuss its meaning.

Definition 7.2 (Sequential group action)**.** We say that the action of the sequence of groups $(G_n)_{n\in\mathbb{N}}$ on $(\mathcal{X}^n)_{n\in\mathbb{N}}$ is sequential if the following conditions hold.

(i) The sequence $(G_n)_{n\in\mathbb{N}}$ is ordered by inclusion: for each n, there is an inclusion map $\imath_{n+1} : G_n \to G_{n+1}$ such that $\imath_{n+1}$ is a continuous group isomorphism between G_n and its image, and the image of G_n under $\imath_{n+1}$ is closed in G_{n+1}.

(ii) For all $g_n \in G_n$ and all $x^{n+1} \in \mathcal{X}^{n+1}$, $\text{proj}_{\mathcal{X}^n}(\imath_{n+1}(g_n)x^{n+1}) = g_n(\text{proj}_{\mathcal{X}^n}(x^{n+1}))$, where $\text{proj}_{\mathcal{X}^n}$ is the canonical projection map $\text{proj}_{\mathcal{X}^n} : \mathcal{X}^{n+1} \to \mathcal{X}^n$ given by $\text{proj}_{\mathcal{X}^n}(x_1, \ldots, x_n, x_{n+1}) = (x_1, \ldots, x_n)$.

(iii) Let $n \geq 1$, $g_n \in G_n$, and $g_{n+1} \in G_{n+1}$. For $x^{n+1} = (x_1, \ldots, x_{n+1}) \in \mathcal{X}^{n+1}$, denote $(x^{n+1})_{n+1} = x_{n+1}$. Then, $g_{n+1} = \imath_{n+1}(g_n)$ if and only if, for all $x^{n+1} \in \mathcal{X}^{n+1}$, $(g_{n+1}x^{n+1})_{n+1} = x_{n+1}$.

In Definition 7.2, item (i) gives an ordering of the sequence of groups by inclusion, (ii) ensures that this inclusion does not change the action of the groups on past data, and (iii) implies that the groups do not act on future data. As a result, invariance of X^{n-1} under G_{n-1} is implied by invariance of X^n under G_n and the individual statements of invariance in (7.1) for each n do not contradict each other. The instances of (7.1) discusssed in Example 7.1 satisfy this assumption; a simpler situation where this is satisfied is given in the next example.

Example 7.2 (Within-batch invariance)**.** Perhaps the simplest example is when, for each n, G_n has a product structure and acts on $\mathcal{X}^n$ componentwise. This is when

$$G_n = H_1 \times H_2 \times \cdots \times H_n$$

for some sequence of topological groups $(H_n)_{n\in\mathbb{N}}$, each H_n acting continuously on $\mathcal{X}$ by $(h, x) \mapsto hx$ and $(g_n, X^n) \mapsto (h_1X_1, \dots, h_nX_n)$ for each $g_n = (h_1, \dots, h_n) \in G_n$. This covers the setting where batches of data are observed sequentially and the interest is in testing group invariance within each batch. For example, assume that $\mathcal{X} = \mathbb{R}^k$, each H_i is a fixed group H acting on $\mathbb{R}^k$, and data $X_1^k, X_2^k, \dots$ are assumed to be i.i.d. copies of a random variable X^k. Then (7.1) becomes the problem of testing sequentially whether $X^k \overset{\mathcal{D}}{=} hX^k$ for all $h \in H$, that is, whether the distribution of X^k is H-invariant. Koning (2023) treats this batch-by-batch setting for general compact groups.

In a addition to this example, sequential group actions also include more complicated situations where there is "cross-action" between the different data points, as in Example 7.1 (corresponding to exchangeability and sphericity). The details for the case of testing rotational symmetry are given in Section 7.5.2.

We now show that, under the assumption that the group action is sequential, the null hypothesis of invariance is an online compression model. The latter are models for computing online summaries, or compressed representations, of the observed data. When the data is generated by an online compression model, the techniques developed for conformal prediction can be used to construct a sequence of statistics that has a Uniform($[0, 1]$) distribution under the null hypothesis. These statistics can be used to build a conformal (test) martingale, as we will discuss in Section 7.4. Vovk et al. define online compression models in abstract terms; we use a simplified definition here.

Definition 7.3 (Online compression model, Vovk et al. (2005))**.** An online compression model on $\mathcal{X}$ is a 3-tuple of sequences $((\sigma_n)_{n\in\mathbb{N}}, (F_n)_{n\in\mathbb{N}}, (Q_n)_{n\in\mathbb{N}})$, where

1. $(\sigma_n)_{n\in\mathbb{N}}$ is a sequence of statistics $\sigma_n = \sigma_n(X^n)$; we call σ_n a summary of X^n,
2. $(F_n)_{n\in\mathbb{N}}$ is a sequence of functions such that $F_n(\sigma_{n-1}, X_n) = \sigma_n$,
3. $(Q_n)_{n\in\mathbb{N}}$ is a sequence of conditional distributions for (σ_{n-1}, X_n) given σ_n.

To show how sequential group invariance defines an online compression model, we first recall some group theory. First, the orbit G_nX^n of X^n under the action of G_n is the set of all values that are reached by the action of G_n on X^n, i.e., $G_nX^n =$

$\{gX^n : g \in G_n\}$. In order to identify each orbit, we pick a single element of $\mathcal{X}^n$ in each orbit—an orbit representative—and consider the map $\gamma_n : \mathcal{X}^n \to \mathcal{X}^n$ that takes each X^n to its orbit representative. We call γ_n an orbit selector (see Section 7.5 for examples), and we assume that it is measurable. Such measurable orbit selectors exist under weak regularity conditions on $\mathcal{X}^n$ and G_n (see Bondar, 1976, Theorem 2) that hold in all the examples of this chapter. Furthermore, because G_n is a compact group, there exists a unique G_n-invariant probability distribution μ_n, called the Haar (probability) measure (Bourbaki, 2004, Chapter VII). The Haar measure plays the role of a uniform probability distribution on compact groups. Finally, it is a fact that the data is uniformly distributed on its orbit conditionally on the orbit where it lays; formally, $X^n \mid \gamma_n(X^n) \overset{\mathcal{D}}{=} U\gamma_n(X^n) \mid \gamma_n(X^n)$, where $U \sim \mu_n$ independently of X (Eaton, 1989, Theorem 4.4).

We now show that if a sequence of groups $(G_n)_{n\in\mathbb{N}}$ acts sequentially on the data, any distribution that is invariant under the action of said sequence defines an online compression model. We use the orbit representative as summary statistic, i.e., we use $\sigma_n = \gamma_n(X^n)$. Then, since the distribution of the data is $(G_n)_{n\in\mathbb{N}}$-invariant, the sequence of conditional distributions of (σ_{n-1}, X_{n-1}) given σ_n is uniform over the orbits as remarked earlier. In this way, we fix σ_n and Q_n in Definition 7.3. Furthermore, the next proposition shows that, for sequential group actions, σ_n can be computed as a function of σ_{n-1} and X_n. The proof of this proposition can be found in Appendix E.1 and it uses crucially the assumption that the group action is sequential. This discussion and the following proposition prove that a sequential group-invariant model indeed defines an online compression model, as we state in Corollary 7.5.

Proposition 7.4. *If the action of $(G_n)_{n\in\mathbb{N}}$ on $(\mathcal{X}^n)_{n\in\mathbb{N}}$ is sequential, then there exists a sequence $(F_n)_{n\in\mathbb{N}}$ of measurable functions $F_n : \mathcal{X}^{n-1} \times \mathcal{X} \to \mathcal{X}^n$ such that $F_n(\gamma_{n-1}(X^{n-1}), X_n) = \gamma_n(X^n)$ and $F_n(\,\cdot\,, X_n)$ is a one-to-one function of $\gamma_{n-1}(X^{n-1})$.*

Corollary 7.5. *Assume that the action of $(G_n)_{n\in\mathbb{N}}$ on $(\mathcal{X}^n)_{n\in\mathbb{N}}$ is sequential, let $\widetilde{\mu}_n$ be the uniform distribution on G_nX^n induced by the Haar measure μ_n on G_n, and let $(F_n)_{n\in\mathbb{N}}$ be as guaranteed by Proposition 7.4. Then the tuple*

$$((\gamma_n(X^n))_{n\in\mathbb{N}}, (F_n)_{n\in\mathbb{N}}, (\widetilde{\mu}_n)_{n\in\mathbb{N}}),$$

defines an online compression model on $\mathcal{X}$.

7.4 Testing Group Invariance With Conformal Martingales

We now construct test martingales for the null hypothesis of distributional symmetry in (7.1) any time that a sequence of groups $(G_n)_{n\in\mathbb{N}}$ acts sequentially on the data $(X_n)_{n\in\mathbb{N}}$. To this end, the invariant structure of the null hypothesis $\mathcal{H}_0$ is used in tandem with conformal prediction to build a sequence of independent random variables $(R_n)_{n\in\mathbb{N}}$ with the following three properties:

1. The sequence $(R_n)_{n\in\mathbb{N}}$ is adapted to the data sequence with external randomization $(X_n, \theta_n)_{n\in\mathbb{N}}$, that is, for each $n \in \mathbb{N}$, $R_n = R_n(X_n, \theta_n)$.
2. Under any element of the null hypothesis $\mathcal{H}_0$ from (7.1), $(R_n)_{n\in\mathbb{N}}$ is a sequence of independent and identically distributed Uniform$[0, 1]$ random variables.
3. The distribution of $(R_n)_{n\in\mathbb{N}}$ is not uniform when departures from symmetry are present in the data.

The construction of these random variables is the subject of Section 7.4.1—additional definitions are needed—and their optimality is the subject of Section 7.4.2. In order to guide intuition, Example 7.3 shows a first example for testing exchangeability, which has previously also been studied by Vovk et al. (2005) and Fedorova et al. (2012). They call the statistics $R_1, R_2, \ldots$ p-values owing to their uniformity. We opt against that terminology here, because typically only small p-values are interpreted evidence against the null hypothesis. However, in the context of testing for symmetry, it is any deviation from uniformity that we interpret as evidence against the null hypothesis. For reasons that will become apparent soon, we call $R_1, R_2, \ldots$ (smoothed) orbit ranks (see Definition 7.7).

With the sequence $(R_n)_{n\in\mathbb{N}}$ at hand, test martingales against distributional invariance are built by testing against the uniformity of $(R_n)_{n\in\mathbb{N}}$. Indeed, any time that $(f_n)_{n\in\mathbb{N}}$ is a sequence of functions $f_i : [0, 1] \to \mathbb{R}$ such that $\int f_i(r)\mathrm{d}r = 1$, the process $(M_n)_{n\in\mathbb{N}}$ given by

$$M_n := \prod_{i\leq n} f_i(R_i) \tag{7.2}$$

is a test martingale for $\mathcal{H}_0$ with respect to $\mathbb{F}$, where $\mathbb{F} = (\sigma(R^n))_{n\in\mathbb{N}}$ and $\sigma(R^n)$ is the σ-algebra generated by R^n. This follows from the fact that $\mathbf{E}_Q\left[M_n \mid \sigma(R^{n-1})\right] = M_{n-1} \int f_n(r)\mathrm{d}r = M_{n-1}$, where we leverage independence and uniformity. The functions $(f_n)_{n\in\mathbb{N}}$ are known as calibrators (Vovk and Wang, 2021). They can be taken to be any sequence of predictable estimators of the density of $R_1, R_2, \ldots$ (Fedorova et al., 2012), so that the test martingale is expected to grow if the true distribution of the orbit ranks is not uniform, i.e., the null hypothesis is violated. The optimality of this procedure is discussed in Section 7.4.2.

Example 7.3 (Sequential Ranks)**.** Consider the case of testing exchangeability as discussed in Example 7.1, that is, the case when $\mathcal{X} = \mathbb{R}$ and $G_n = S(n)$. For each n, define the random variables $\widetilde{R}_n = \sum_{i \le n} \mathbf{1}\{X_i \le X_n\}$—the rank of X_n among $X_1, \ldots, X_n$. The random variables $\widetilde{R}_1, \widetilde{R}_2, \ldots$ are called sequential ranks (Malov, 1996). It is a classic observation that each $\widetilde{R}_n$ is uniformly distributed on $\{1, \ldots, n\}$, and that $(\widetilde{R}_n)_{n\in\mathbb{N}}$ is a sequence of independent random variables (Rényi, 1962). After rescaling and adding external randomization, a sequence of random variables $(R_n)_{n\in\mathbb{N}}$ can be built from $(\widetilde{R}_n)_{n\in\mathbb{N}}$ such that $(R_n)_{n\in\mathbb{N}}$ satisfies items 1, 2 and 3 at the start of this section. Furthermore, if we denote the uniform measure on $S(n)$ by μ_n, then $\widetilde{R}_n$ can also be obtained from $n^{-1}\widetilde{R}_n = \mu_n\{g : (gX_n)_n \le X_n\}$. While this rewriting may seem esoteric at this point, it turns out to be the correct point of view for generalization.

7.4.1 Conformal Prediction Under Invariance

In general, the statistics R_n will be designed to measure how strange the observations X^n are in contrast to what would be expected under distributional invariance. To this end, the values of X^n are compared to those in the orbit of X^n under the action of G_n. In order to measure the "strangeness" of the observations in their orbit, we use an adaptation of the conformity measures introduced by Vovk et al. (2005).

Definition 7.6 (Conformity measure of invariance)**.** We say that the function $A : \mathcal{X} \times \bigcup_{n=1}^{\infty} \mathcal{X}^n \to \mathbb{R}$ is a conformity measure of invariance if the following holds: if there are $X^n, X'^n \in \mathcal{X}^n$, such that $A(X_i, \gamma_n(X^n)) = A(X'_i, \gamma_n(X'^n))$ for all $i \in \{1, \ldots, n\}$, then, for all $g \in G_n$, we also have that $A((gX^n)_i, \gamma_n(X^n)) = A((gX'^n)_i, \gamma_n(X'^n))$ for all $i \in \{1, \ldots, n\}$.

The group-related condition on A that appears in Definition 7.6 is an addition to that of Vovk et al. (2005); it ensures that the action of G_n on $\mathcal{X}^n$ induces an action on the conformity measure. The intuition of the definition is that when A is properly chosen, $A(X_n, \gamma_n(X^n))$ is a numerical score that indicates how similar X_n is to the other values in its orbit. Therefore, the statistic $\alpha_n = A(X_n, \gamma_n(X^n))$ is called a conformity score. The easiest example is when $\mathcal{X}^n \subseteq \mathbb{R}^n$ because then A defined by $A(X_n, \gamma_n(X^n)) = X_n$ is a conformity measure of invariance—this is the case in Example 7.3. However, perhaps a more intuitive choice would be $A(X_n, \gamma_n(X^n)) = |X_n - \int_G g\gamma_n(X^n)\, \mathrm{d}\mu_n(g)|^{-1}$, since this quantity is large whenever X_n is close to the average value within the orbit, which is given by $\int_G g\gamma_n(X^n)\, \mathrm{d}\mu_n(g)$. For a more involved example, consider the case where the data points are given by $X_i = (Y_i, Z_i)$ for some outcome $Y_i \in \mathbb{R}$ and a covariate $Z_i \in \mathbb{R}$. Then one might consider $A(X_i, \gamma_n(X^n)) = |Y_i - \hat{Y}_i|^{-1}$, where $\hat{Y}_i$ is the prediction of some regression method that was trained on the orbit of X_i. In this case, the intuition is that, if the label is very close to the prediction that is made using all of the values in the same orbit, then X_i must have been very typical of the orbit. For a more detailed discussion on different conformity measures, we refer to Fontana et al. (2023).

Since the scale of the conformity scores is arbitrary—they can be scaled at will—, only comparisons between them are meaningful. Therefore, similar to what happened in Example 7.3, we will rank the observed value of the conformity score α_n among all its possible values on the orbit of the data. To this end, we obtain the distribution of the conformity scores under the null hypothesis using the assumed distributional invariance. Indeed, as discussed in Section 7.3, the distribution of X^n conditional on $\gamma(X^n)$ is uniform on its orbit. This idea gives rise to the (smoothed) orbit ranks $(R_n)_{n\in\mathbb{N}}$ in the next definition.

Definition 7.7 (Smoothed Orbit Ranks)**.** Fix $n \in \mathbb{N}$, let A be a conformity measure, and let $\alpha_n = A(X_n, \gamma_n(X^n))$ be the associated conformity score. We call R_n, defined by

$$\begin{aligned} R_n =& \mu_n(\{g \in G_n : A((gX^n)_n, \gamma_n(X^n))_n < \alpha_n\}) + \\ & \theta_n \mu_n(\{g \in G_n : A((gX^n)_n, \gamma_n(X^n)) = \alpha_n\}), \end{aligned} \tag{7.3}$$

a (smoothed) orbit rank, where μ_n denotes the Haar probability measure on G_n and $\theta_n \sim \mathrm{Uniform}[0,1]$ is independent of the data X^n.

The simplest case is when the group G_n is finite of size k and $A(X_i, \gamma_n(X^n)) = X_i$. In that case, μ_n is the discrete uniform distribution on G_n and $R_n = \frac{1}{k}\#\{g \in G_n : (gX^n)_n < X_n\} + \frac{\theta_n}{k}\#\{g \in G_n : (gX^n)_n = X_n\}$.

An important intuition is that the statistic R_n is the CDF of the distribution of α_n conditional on $\gamma_n(X^n)$ evaluated in α_n (with added randomization) under $\mathcal{H}_0$. It follows that if said CDF is continuous, smoothing plays no role in (7.3) and $R_n \perp \theta_n$. It also follows—and this is shown in Theorem 7.8—that each R_n is uniformly distributed on $[0, 1]$. Vovk et al. (2005, Theorem 11.2) show that, if the data is generated by an online compression model, and $\theta_1, \theta_2, \dots$ are independent, then $R_1, R_2, \dots$ are also independent. Since Corollary 7.5 shows that a sequential group invariance structure defines an online compression model, it follows that the smoothed orbit ranks form an i.i.d. uniform sequence under the null hypothesis. This is stated in the next theorem, for which we provide a direct proof in Appendix E.1 for completeness .

Theorem 7.8. *Suppose that the action of $(G_n)_{n\in\mathbb{N}}$ on $(\mathcal{X}^n)_{n\in\mathbb{N}}$ is sequential, that $(X_n)_{n\in\mathbb{N}}$ is generated by an element of $\mathcal{H}_0$, and that $\theta_1, \theta_2, \dots$ are independent. Then $R^n \perp \gamma_n(X^n)$ for each n and $(R_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d.* Uniform$[0,1]$ *random variables.*

7.4.2 Optimality

We now show that any martingale based on the smoothed orbit ranks as in (7.2) is a likelihood ratio process and that it is log-optimal against the implicit alternative for which it is built. To this end, let P be a distribution such that for all n, conditionally on R^{n-1}, R_n has density f_n (with respect to the Lebesgue measure). Technically, the conditional density of R_n is not defined only by P, but also by the external randomization. To make this explicit in the following, we will use $\widetilde{P}$ to denote P with external randomization added, that is, $\widetilde{P} = P \times \mathcal{U}^\infty$, where $\mathcal{U}^\infty$ is the uniform distribution on $[0,1]^\infty$. Analogously, for each $Q \in \mathcal{H}_0$, define $\widetilde{Q} = Q \times \mathcal{U}^\infty$.

The discussion below (7.2) shows that $M_n = \prod_{i\leq n} f_i(R_i)$ is a test martingale. In fact, M_n is the likelihood ratio for the orbit ranks R^n between $\widetilde{P}$ and $\widetilde{Q}$, since the distribution of R^n under $\widetilde{Q}$ equals the uniform distribution for any $Q \in \mathcal{H}_0$ by Theorem 7.8. Surprisingly, if $\widetilde{P}$ is such that $R_n \perp \gamma_n(X^n)$, then M_n is also the likelihood ratio for the full data X^n between P and an appropriately chosen distribution $Q^* \in \mathcal{H}_0$, as shown in the following proposition. For the sake of brevity, the action of $(G_n)_{n\in\mathbb{N}}$ on $(\mathcal{X}^n)_{n\in\mathbb{N}}$ is assumed to be sequential throughout.

Proposition 7.9. *Suppose that $A(\,\cdot\,, \gamma_n(X^n))$ is a one-to-one function for each $n \in \mathbb{N}$, suppose that P is any distribution under which $R_n \perp \gamma_n(X^n)$ for each n, and let f_i denote the conditional density of R_i given R^{i-1} under P. Let $M_n = \prod_{i\leq n} f_i(R_i)$. Then, for $Q \in \mathcal{H}_0$,*

$$\widetilde{Q}\left(M_n = \frac{\mathrm{d}P}{\mathrm{d}Q^*}(X^n)\right) = 1, \tag{7.4}$$

where Q^* denotes the distribution under which the marginal of $\gamma_n(X^n)$ coincides with that under P, and such that $X^n \mid \gamma_n(X^n) \overset{\mathcal{D}}{=} U\gamma_n(X^n) \mid \gamma_n(X^n)$, where $U \sim \mu_n$ independently from $\gamma_n(X^n)$.

The distribution Q^* can be thought of as a symmetrization of P, since the marginal of $\gamma_n(X^n)$ is the same, but the distribution conditional on $\gamma_n(X^n)$ is defined by symmetry. Proposition 7.9 therefore shows that, if the orbit ranks are independent of the orbit selectors under P, then $(M_n)_{n\in\mathbb{N}}$ is the likelihood ratio process between P and a symmetrization thereof. The next theorem uses this representation to show the log-optimality of $(M_n)_{n\in\mathbb{N}}$. Its proof follows that of Theorem 12 of Koolen and Grünwald (2022).

Theorem 7.10. *Assume that $A(\,\cdot\,,\gamma_n(X^n))$ is one-to-one for all $n\in\mathbb{N}$ and let P be such that, under P, the distribution of $X^n \mid \gamma_n(X^n)$ is absolutely continuous with respect to the uniform distribution. Denote f_i for the density of $R_i \mid R^{i-1}$ under $\widetilde{P}$ and let $M_n = \prod_{i\le n} f_i(R_i)$. Let τ be any stopping time and $(E_n)_{n\in\mathbb{N}}$ any e-process for $\mathcal{H}_0$, both with respect to $\mathbb{F}$—the filtration generated by the smoothed ranks. Then it holds that*

$$\mathbf{E}_{\widetilde{P}}\left[\ln M_\tau\right] = \mathbf{E}_{\widetilde{P}}\left[\ln \prod_{i=1}^{\tau} f_i(R_i)\right] \ge \mathbf{E}_{\widetilde{P}}\left[\ln E_\tau\right]. \tag{7.5}$$

Moreover, if $\widetilde{P}$ is such that $R^n \perp \gamma_n(X^n)$ for all n, then for any e-process E' for $\mathcal{H}_0$ w.r.t. $(\sigma(X^n,\theta^n))_{n\in\mathbb{N}}$—the full-data filtration—, it also holds that

$$\mathbf{E}_{\widetilde{P}}\left[\ln M_\tau\right] \ge \mathbf{E}_{\widetilde{P}}\left[\ln E'_\tau\right]. \tag{7.6}$$

The first part of Theorem 7.10, Equation (7.5), establishes that, under some assumptions on P, $(M_n)_{n\in\mathbb{N}}$ is log-optimal for testing group invariance among all e-processes defined only on the orbit ranks. Moreover, the second part of Theorem 7.10, Equation (7.6), states that if the orbit ranks are also independent of the orbit selector under P, then $(M_n)_{n\in\mathbb{N}}$ is log-optimal for testing group invariance among all e-processes defined on the full data. The additional assumption of independence between R^n and $\gamma_n(X^n)$ is necessary for (7.6) to hold: if $\widetilde{P}$ is a distribution under which $R_1,\dots,R_n \not\perp \gamma_n(x^n)$, then the conformal martingale is not in general a likelihood ratio as in (7.4). For the deterministic stopping time $\tau=n$, the log-optimal statistic is instead given by $S_n = \prod_{i=1}^n f_n(R_1,\dots,R_n \mid \gamma_n(X^n))$, as it can be written as a likelihood ratio (see also Grünwald et al., 2024; Koning, 2023). However, the sequence $(S_n)_{n\in\mathbb{N}}$ does not necessarily define a test martingale or e-process, so that it might not be possible to use it in the construction of an anytime-valid test. Using tests based on the sequential ranks circumvents this issue for such alternatives.

The optimality of M_n in Theorem 7.10 is contingent on oracle knowledge of the true distributions $f_1, f_2, \dots$, which are unknown in practice. To counter this, past data can be used to sequentially estimate the true density. This idea has previously been applied for testing exchangeability (Vovk et al., 2005; Fedorova et al., 2012). More precisely, for each n, let $\hat{f}_n$ be an estimator of f_n based on R^{n-1}, and consider the martingale defined by $\prod_{i=1}^{n} \hat{f}_i(R_i)$. In general, this is suboptimal with respect to an oracle that knows the true density. However, in the case that there exists a density f such that $f_i \equiv f$ for all i, i.e. data are i.i.d. under P, there is limited loss asymptotically if $\hat{f}_i$ is a good estimator of f. In order to judge if an estimator is good for the task at hand, consider the difference in expected growth per outcome for fixed n, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\widetilde{P}}[\log f(R_i) - \log \hat{f}_i(R_i)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\widetilde{P}}[\mathrm{KL}(f \| \hat{f}_i)], \tag{7.7}$$

where $\mathrm{KL}(f\|\hat{g}) = \int_0^1 f(r)\log(f(r)/g(r))\mathrm{d}r$ denotes the Kullback-Leibler divergence whenever f is absolutely continuous with respect to g, and the expectation on the right-hand side of (7.7) is over past data (on which $\hat{f}_i$ depends). If (7.7) tends to zero as n grows large, the expected growth per outcome converges to that of the log-optimal test martingale. This motivates the use of density estimation algorithms for which this always happens. Under stringent assumptions—for example, if the density f belongs to an exponential family—sequential Bayesian-update-type algorithms are known to guarantee that (7.7) converges to zero (Kotłowski and Grünwald, 2011). Under weaker assumptions, specialized algorithms exist with the same guarantees (Haussler and Opper, 1997; Cesa-Bianchi and Lugosi, 2001; Grünwald and Mehta, 2019).

7.5 Applications and Extension

In this section, we discuss applications and an extension of the theory developed in the previous sections.

7.5.1 Sign-Invariant Exchangeability

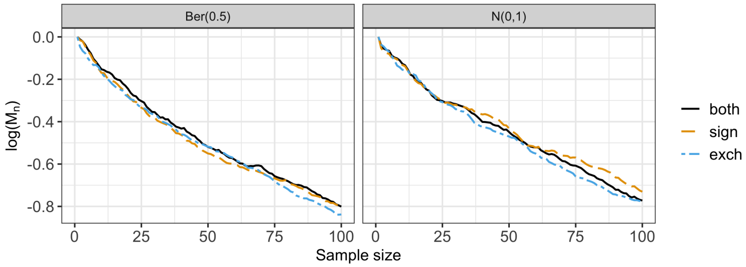
In this subsection, we consider testing for sign-invariant exchangeability (Berman, 1965; Fraiman et al., 2024) with the purpose of illustrating our method on a concrete, basic example, and show its performance through numeric simulation. Real-valued data are sign-invariant if $(X_1, \ldots, X_n) \stackrel{\mathcal{D}}{=} (\epsilon_1 X_1, \ldots, \epsilon_n X_n)$ for all signs $(\epsilon_1, \ldots, \epsilon_n) \in \{-1, 1\}^n$. We consider $\{-1, 1\}^n$ as a group with componentwise multiplication as operation. Data are sign-invariant exchangeable if they are both sign-invariant and exchangeable. That is, if $(X_1, \ldots, X_n) \stackrel{\mathcal{D}}{=} (\epsilon_1 X_{\pi(1)}, \ldots, \epsilon_n X_{\pi(n)})$ for all signs $(\epsilon_1, \ldots, \epsilon_n) \in$

$\{-1,1\}^n$ and all permutations $\pi \in S(n)$. The null hypothesis of sign-invariant exchangeability is therefore equivalent to the distributional invariance of X^n under the action of $G_n = \{-1,1\}^n \times S(n)$. The orbit of X^n under G_n is given by the set $\{(\epsilon_1 X_{\pi(1)}, \ldots, \epsilon_n X_{\pi(n)}) : (\epsilon_1, \ldots, \epsilon_n) \in \{-1,1\}^n, \pi \in S(n)\}$. Since G_n is finite, the Haar measure is the discrete uniform distribution. Furthermore, because the data are assumed to be real, we can take $A(X_n, \gamma_n(X^n)) = X_n$. Let $X'_1, \ldots, X'_{2n}$ be given, for $i = 1, \ldots, n$, by $X'_i = X_i$ and $X'_{n+i} = -X_i$. The smoothed orbit rank in (7.3) becomes

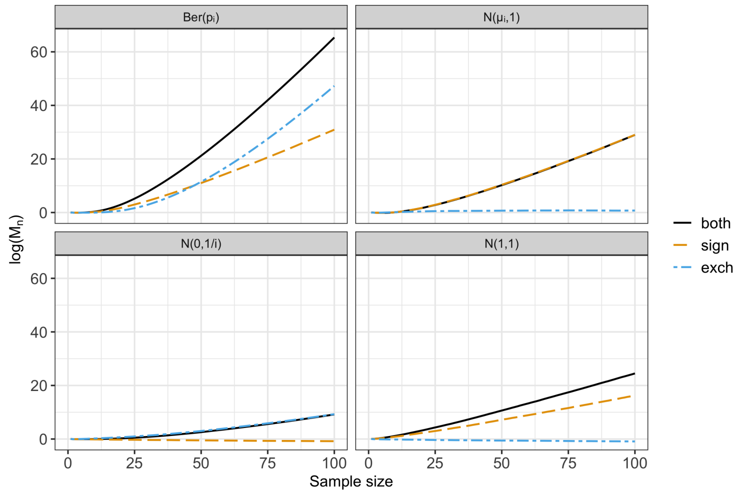
$$R_n = \frac{\#\{i \leq 2n : X'_i < X_n\}}{2n} + \theta_n \frac{\#\{i \leq 2n : X'_i = X_n\}}{2n}. \tag{7.8}$$

These statistics can be computed upon observing the data, and standard density estimation algorithms can used to estimate their density. Following Section 7.4, for each n we use $R_1, \ldots, R_{n-1}$ to build an estimate $\hat{f}_n$ of the density and use $M_n = \prod_{i \leq n} \hat{f}_i(R_i)$ as a test martingale.

We investigate how martingales obtained in this manner behave through simulations. For these experiments, we used the R language (R Core Team, 2022). The density estimation was performed using the kernel density estimation that is implemented in the `Stats` package. However, standard kernel density estimation can lead to poor performance around the boundaries. This is because these algorithms are designed to estimate densities supported on $\mathbb{R}$ and not just on $[0,1]$. Following Fedorova et al. (2012), a solution to this problem is (in the case of testing exchangeability) to reflect the sequence of orbit ranks to the left from zero and to the right from one. Then, the estimate is computed using the extended sample $\cup_{i=1}^{n} \{-R_i, R_i, 2 - R_i\}$. Finally, the estimated density is set to zero outside of the unit interval and then normalized. We have used this same procedure here. Furthermore, for the sake of comparison, we also include the conformal martingale that would be obtained if testing either for exchangeability exclusively or for sign-invariance. The results are shown in Figure 7.1. We see that, if data are sign-invariant exchangeable—i.i.d. Rademacher or i.i.d. Normal$(0,1)$ in our experiments—, the conformal martingales are indeed martingales and do not take large values, as expected based on the discussion in Section 7.2. Under the alternative, the statistic M_n is no longer a martingale, and it does grow. However, the methods that test for only one of the two symmetries (either exchangeability or sign invariance separately) do not detect alternatives for which that particular symmetry is not violated, but the other is (see Figure 7.1b). On the other hand, the conformal martingale based on R_n as described in (7.8) detects all of the alternatives. In fact, for the alternative where each $X_i \in \{-1,1\}$ and $X_i = 1$ with probability $p_i = 1 - 1/i$



(a) Under null models: $X_i = \pm 1$ w.p. 0.5 and $X_i \sim N(0,1)$.



(b) Statistics under alternative models (all independent). Upper left corner: $X_i = 1$ w.p. $p_i = 1 - 1/i$ and $X_i = 1$ with probability $1 - p_i$. Upper right corner: $X_i \sim N(i/10, 1)$. Lower left corner: $X_i \sim N(0, 1/i)$. Lower right corner: $X_i \sim N(1,1)$.

Figure 7.1: The logarithm of the conformal martingale against sample size for three different methods: testing for both sign-invariance and exchangeability, or only one of the two. A test built against sign-invariance and exchangeability can detect the absence of either of those two invariances while a test that is built to detect only one of then cannot achieve the same goal (see Section 7.5.1). Data are independent under all considered models. The results were averaged over 500 repetitions.

and $X_i = -1$ with probability $1 - p_i$ indepently, the corresponding test martingale is even log-optimal among all e-processes. This is due to the fact that, regardless of the observed data, the orbit of X_i is always the set $\{-1, 1\}$. Therefore, the orbit selector can be chosen to be $\gamma_n(X^n) = (1, \ldots, 1)$ independently of the data, such that $R_n \perp \gamma_n(X^n)$. The log-optimality then follows from Theorem 7.10.

7.5.2 The Orthogonal Group and Linear Models

Consider testing whether the data we observe are drawn from a spherically symmetric distribution, i.e., $\mathcal{X} = \mathbb{R}$ and $G_n = \mathrm{O}(n)$, where $\mathrm{O}(n)$ is the orthogonal group in dimension n. Testing for spherical symmetry is equivalent to testing whether the data are generated by a zero-mean Gaussian distribution. Indeed, any distribution on $\mathbb{R}^\infty$ for which the marginal of the first n coordinates is spherically symmetric for any n, is a mixture of i.i.d. zero-mean Gaussian distributions (Bernardo and Smith, 2009, Proposition 4.4). It follows that any process that is a supermartingale under all zero-mean Gaussian distributions is also a supermartingale under spherical symmetry and vice-versa. This implies that, for the purpose of testing with test (super)martingales, the two hypotheses are equivalent. We show how this fits in our setting, and defer the application to regression to Appendix E.2.

We now check that testing spherical symmetry fits in our setting, i.e., that Definition 7.2 is fulfilled. Consider the inclusion of $\mathrm{O}(n)$ in $\mathrm{O}(n+1)$ given by

$$\imath_{n+1}(O_n) = \begin{pmatrix} O_n & 0 \\ 0 & 1 \end{pmatrix}$$

for each $O_n \in \mathrm{O}(n)$. Using the canonical projections in $\mathbb{R}^n$, Definition 7.2 is readily checked. Since the data are real, we can consider the simple measure of conformity $A(X_n, \gamma_n(X^n)) = X_n$. An orbit selector is given by $\gamma_n(X^n) = \|X^n\| e_1$, where e_1 is the unit vector $e_1 = (1, 0, \ldots, 0)$. For simplicity, we assume that the distribution of X^n has a density with respect to the Lebesgue measure for each n, so that $R_n = \mu_n(\{O_n \in \mathrm{O}(n) : (O_n X^n)_n < X_n\})$—no external randomization is needed. Rather than thinking of μ_n as a measure on $O(n)$, one can think of it as the uniform measure on $S^{n-1}(\|X^n\|)$. This way, R_n can be recognized to be the relative surface area of the hyper-spherical cap with co-latitude angle $\varphi_n = \pi - \cos^{-1}(X_n/\|X^n\|)$. Li (2010) shows that an explicit expression for this area is given by

$$R_n = \begin{cases} 1 - \frac{1}{2} I_{\sin^2(\pi - \varphi_n)}\left(\frac{n-1}{2}, \frac{1}{2}\right) & \text{if } \varphi_n > \frac{\pi}{2}, \\ \frac{1}{2} I_{\sin^2(\varphi_n)}\left(\frac{n-1}{2}, \frac{1}{2}\right) & \text{else,} \end{cases} \tag{7.9}$$

where $I_x(a, b)$ denotes the regularized beta function, $I_x(a,b) = \frac{B(x,a,b)}{B(1,a,b)}$ for $B(x, a, b) = \int_0^x t^{a-1}(1-t)^{b-1}$ for $0 \le x \le 1$.

Note that $\varphi_n > \frac{\pi}{2}$ if and only if $X_n > 0$ and that $\sin^2(\varphi_n) = 1 - \frac{X_n^2}{\|X^n\|^2}$, so that (7.9) equals the CDF of the t-distribution with $n-1$ degrees of freedom evaluated in $t = \sqrt{n-1}X_n/\|X^{n-1}\|$. If $X^n \sim \mathcal{N}(0, \sigma^2 I_n)$, then t is the ratio of a normally distributed random variable and an independent chi-squared-distributed random variable. Therefore, t has a t-distribution with $n-1$ degrees of freedom. The test thus obtained is a type of sequential t-test that has, to the best of our knowledge, not been considered previously.

This example can be extended to testing for centered spherical symmetry, i.e., whether $X^n = \mu \mathbf{1}_n + \epsilon^n$, where $\mathbf{1}_n$ is the all-ones n-vector, $\mu \in \mathbb{R}$ and ϵ^n is spherically symmetric for every $n \in \mathbb{N}$. By similar reasoning as above, this is equivalent to testing whether the data is i.i.d. Gaussian with any mean/variance. Even more, by considering different isotropy groups, one can also cover the case where the mean μ is not fixed, but depends on covariates. The ideas needed in that case are similar; we show how deal with the added complexity in Appendix E.2.

7.5.3 Modification for Independence Testing

We now propose a minor modification of the conformal martingales from the previous section that can be used to test for independence. Formally, fix $K \in \mathbb{N}$ and suppose that at each time point $n \in \mathbb{N}$, a K-dimensional vector $X_n = (X_{1,n}, \ldots, X_{K,n}) \in \mathcal{X}^K$ is observed. We are interested in testing the null hypothesis that states that: (1) for each $k = 1, \ldots, K$ and each n the vectors $(X_{k,1}, \ldots, X_{k,n})$ are G_n-invariant, and (2) $(X_{k,1}, \ldots, X_{k,n}) \perp (X_{k',1}, \ldots, X_{k',n})$ for all $k \neq k' \in \{1, \ldots, K\}$. Under this hypothesis, the data is invariant under the sequential action of $(\widetilde{G}_n)_{n\in\mathbb{N}}$ given by $\widetilde{G}_n = G_n^K$, acting on $\mathcal{X}^{K\times n}$ rowwise. That is, the first copy of the group acts on $(X_{1,1}, \ldots, X_{1,n})$, the second on $(X_{2,2}, \ldots, X_{2,n})$, etc. This action is sequential anytime that the action of $(G_n)_{n\in\mathbb{N}}$ is sequential on each of the K data streams.

Based on the discussion above, a first idea to test for invariance under $\widetilde{G}_n$ is to create K test martingales and combine them through multiplication. More specifically, we can treat each of the sequences $(X_{k,n})_{n\in\mathbb{N}}$, $k \in \{1, \ldots, K\}$ as a separate data stream and compute the corresponding statistics in (7.3), leading to K sequences of uniformly distributed random variables $(R_{k,n})_{n\in\mathbb{N}}$. If, for all $n \in \mathbb{N}$ and $k \in \{1, \ldots, K\}$, $f_{k,n}$ is a density on $[0,1]$ then, by independence, the sequence $(M'_n)_{n\in\mathbb{N}}$ defined by $M'_n = \prod_{i=1}^n \prod_{k=1}^K f_{k,i}(R_{k,i})$ is a martingale under the null hypothesis. However, this martingale would not be able to detect alternatives under which the marginals are group invariant, but not independent. This stems from the fact that it only uses that the marginals are uniform under the null, while in fact a stronger claim is true: for each n, the joint distribution of $R_{k,n}$, $k \in \{1, \ldots, K\}$, is uniform on $[0,1]^K$. As a result of this observation, one can choose any sequence of joint density (estimators) $f_1, f_2, \ldots$ on $[0,1]^K$ and create a test martingale by considering $M_n = \prod_{i=1}^n f_i(R_{1,i}, \ldots, R_{K,i})$.

In the case that $K = 2$ and $G_n = S(n)$, this is the procedure that was recently employed by Henzi and Law (2024). They discuss a specific choice of f_n, a histogram density estimator, that is able to detect departures from independence consistently under the stronger assumption that data are i.i.d. One of their key insights is that independence of the data streams not only implies joint uniformity of the sequential ranks in their setting, but that independence and joint uniformity are actually equivalent. This equivalence breaks down if one does not assume that $X_{k,1}, X_{k,2} \ldots$ are i.i.d. for all k. Finding conditions under which the independence of the streams and the joint uniformity of the rank distributions are equivalent so that a histogram density estimator might reliably detect independence in the more general setting is an interesting avenue of research.

7.6 Discussion

We have discussed how the theory of conformal prediction can be applied to test for symmetry of infinite sequences of data. Here we discuss two topics. First, the relationship to noninvariant conformal martingales. Second, whether smoothing is necessary when defining orbit ranks.

7.6.1 Noninvariant Conformal Martingales

Not all online compression models correspond to a compact-group invariant null hypothesis. An interesting example of this phenomenon is when the data are i.i.d. and exponentially distributed. This distribution is invariant under reflections in any 45° line (not necessarily through the origin), but these reflections do not define a compact group and therefore do not fit the setting discussed in this chapter. Nevertheless, the sum of data points is a sufficient statistic for the data, so this model can still be seen as an online compression model with the sum being the summary. More work is needed to find out whether conformal martingales are log-optimal against certain alternatives in such settings.

7.6.2 The Need for Smoothing

In situations when, conditionally on the orbit selector $\gamma_n(X^n)$, the conformity measure $\alpha^n(X^n)$ has a continuous distribution, the smoothing plays no role in (7.7). This is the case for the rotations discussed in Section 7.5.2. In certain other scenarios, smoothing can be avoided as well. Indeed, one can always define nonsmoothed orbit ranks, in opposition to the smoothed ranks R_n from Definition 7.7, by $\widetilde{R}_n := \mu_n(\{g \in G_n : (g\alpha^n)_n \leq \alpha_n\})$. Notice that this nonsmooth version satisfies $\widetilde{R}_n \leq R_n$. For a particular choice of increasing densities $f_1, f_2, \ldots,$ on $[0,1]$—in the sense that $u \mapsto f_i(u)$ is increasing—, we have that the process $\widetilde{M}_n := \prod_{i=1}^n f_i(\widetilde{R}_i)$ is bounded from above by the conformal martingale $M_n = \prod_{i=1}^n f_i(R_i)$. Such a choice of increasing f_i is natural when high values of R_i (or $\widetilde{R}_i$) are associated with departures from the null hypothesis. Then, any sequential test based on an upper threshold on $\widetilde{M}_n$ inherits the anytime-valid type-I error guarantees of M_n—exactly because $\widetilde{M}_n \leq M_n$. This was previously noted by Vovk et al. (2003). However, the process $\widetilde{M}_n$ may not be a martingale itself. Instead, a test martingale can sometimes directly be associated to $\widetilde{R}_n$. For instance, in the setting of Example 7.3 (testing exchangeability), the distribution of $\widetilde{R}_n$ under the null hypothesis is known—it is uniformly distributed on $\{1, \ldots, n\}$. Therefore, we can construct likelihood ratio processes for the sequence of nonsmoothed ranks. Even more, there are parametric alternatives under which the exact distributions of the nonsmoothed ranks can be computed. This is the case for Lehmann alternatives where, under the null, each X_i is assumed to be sampled from some continuous distribution with c.d.f. $F_i(x) = F_0(x)$ for some fixed F_0; under the alternative, $F_i(x) = 1 - (1 - F_0(x))^{\theta_i}$ for some θ_i. From Theorem 7.a.1 of Savage (1956) the distribution of $\widetilde{R}_i$ can be derived, so that the likelihood ratio process of $\widetilde{R}_i$ can be used for testing, thus avoiding external randomization.