How can one test if a binary sequence is exchangeable? Fork-convex hulls, supermartingales, and Snell envelopes

Aaditya Ramdas[1], Johannes Ruf[2], Martin Larsson[3], Wouter M. Koolen[4]

[1] Departments of Statistics and ML, Carnegie Mellon University
[2] Department of Mathematics, London School of Economics
[3] Department of Mathematics, Carnegie Mellon University
[4] Machine Learning Group, CWI Amsterdam

{aramdas,martinl}@andrew.cmu.edu
wmkoolen@cwi.nl, j.ruf@lse.ac.uk

February 1, 2021

Abstract

Suppose we observe an infinite series of coin flips $X_1, X_2, \ldots$, and wish to sequentially test the null that these binary random variables are exchangeable against Markovian alternatives. We utilize a geometric concept called "fork-convexity" (an adapted analog of convexity) that lies at the heart of this problem, and relate it to other concepts like Snell envelopes that are absent in the sequential testing literature. By demonstrating that the alternative lies within the *fork-convex hull* of the null, we prove that any nonnegative supermartingale under the exchangeable null is necessarily also a supermartingale under the alternative, and thus yields a powerless test. We then combine ideas from universal inference (maximum likelihood under the null) and the method of mixtures (Jeffreys' prior over the alternative) to derive a nonnegative process that is upper bounded by a martingale, but is not itself a supermartingale. We show that this process yields safe e-values, which in turn yield sequential level-α tests that are consistent (power one), using regret bounds from universal coding to demonstrate their rate-optimal power. We present ways to extend these results to any finite alphabet and to Markovian alternatives of any order using a "double mixture" approach. We also discuss their power against change point alternatives, and give general approaches based on betting for unstructured or ill-specified alternatives.

Keywords: Anytime-valid sequential inference; betting; calibrator; composite Snell envelope; de Finetti mixing; fork-convexity; Jeffreys' prior; method of mixtures; nonnegative supermartingale; optional stopping; regret bound; safe e-value; testing exchangeability; universal coding.

Contents

1 Safe e-values and nonnegative supermartingales

Suppose we observe a sequence of binary coin flips $X_1, X_2, \cdots$. Consider the problem of testing if our data $(X_t)_{t \in \mathbb{N}}$ is either an exchangeable sequence, or an i.i.d. Bernoulli sequence, and if not, then to stop collecting data and reject the null as soon as possible.

Let the null set $\mathcal{Q}$ consist of all product distributions P^∞, where $P = \text{Ber}(p)$ for some $p \in [0, 1]$. Let $(\mathcal{F}_t)_{t \in \mathbb{N}_0}$ represent the canonical filtration, where $\mathcal{F}_0$ is the trivial sigma algebra and $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$. All martingale statements in this paper will implicitly refer to this canonical filtration. Furthermore, let $\mathcal{T}$ be the set of all stopping times (potentially infinite) with respect to $(\mathcal{F}_t)$. A level α sequential test for this problem is any stopping time $\tau \in \mathcal{T}$ such that

$$\sup_{Q \in \mathcal{Q}} Q(\tau < \infty) \leq \alpha; \tag{1}$$

meaning that with probability $1 - \alpha$, we never stop under the null. Following recent literature on sequential testing [6, 12], we introduce the related notion of a $\mathcal{Q}$-safe e-value, which is a nonnegative sequence of adapted random variables $(E_t)_{t \in \mathbb{N}_0}$ such that

$$\sup_{Q \in \mathcal{Q}} \sup_{\tau \in \mathcal{T}} \mathbb{E}_Q[E_\tau] \leq 1.$$

Above, we interpret $E_\infty := \limsup_{t \to \infty} E_t$ for potentially infinite stopping times. Large e-values encode evidence against the null, and it is easy to check that the stopping time

$$\tau_\alpha := \inf\{t \geq 1 : E_t \geq \tfrac{1}{\alpha}\} \tag{2}$$

results in a level α sequential test by Markov's inequality. More details can be found in Ramdas et al. [12], who also show that the sequence (p_t) defined by $p_t := \inf_{s \geq t} 1/E_s$ is an anytime-valid p-value, meaning:

$$\sup_{Q \in \mathcal{Q}} \sup_{\tau \in \mathcal{T}} Q(p_\tau \leq \alpha) \leq \alpha \tag{3}$$

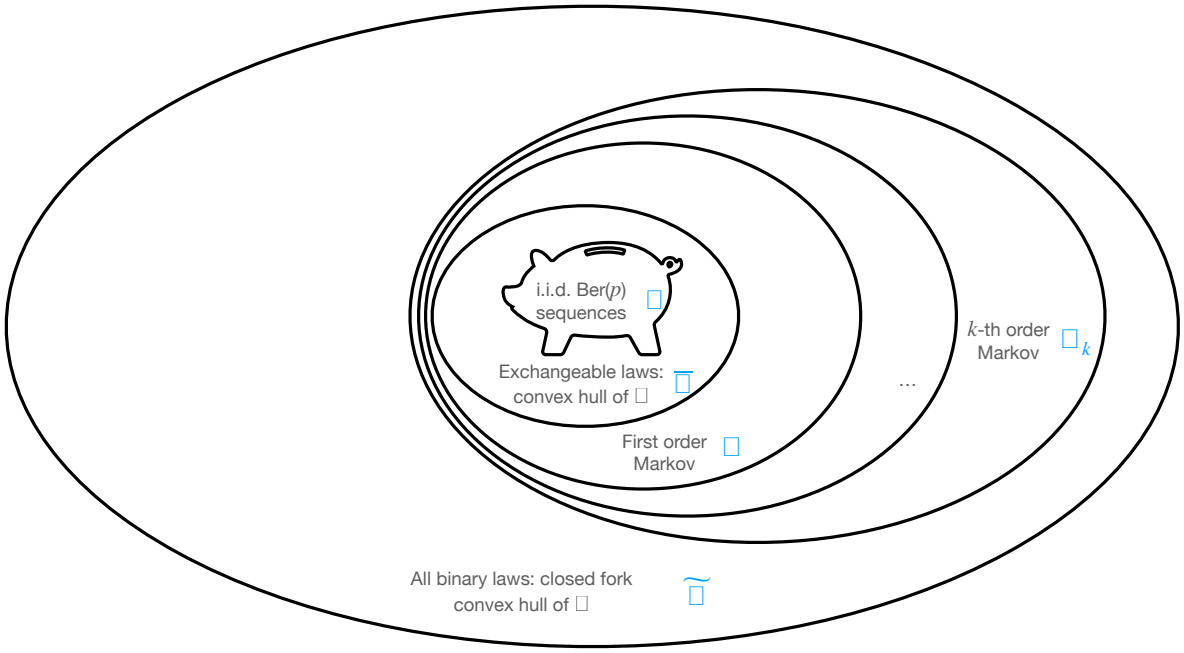

Figure 1: Various classes of distributions over infinite binary sequences encountered in this paper.

for all $\in [0; 1]$. Since such $\mathcal{Q}$-safe e-values result in both sequential tests and anytime-valid p-values, we focus on constructing e-values for the rest of this paper. As a matter of convention, we always use to denote the above stopping time, that is the one that thresholds a safe e-value at level $1=$, while denotes a generic stopping time.

Let $\mathcal{P}$ represent our alternative class, where each $\mathsf{P} \in \mathcal{P}$ represents a first-order Markov process with parameters p_{1j0} and p_{1j1}. Here we abbreviate $p_{0j0} = 1 - p_{1j0}$ and $p_{0j1} = 1 - p_{1j1}$. For simplicity, assume that the first outcome is equally likely to be zero or one. We further assume that the Markov chain is non-absorbing, meaning that $p_{0j1} > 0$ and $p_{1j0} > 0$. Otherwise, with probability half, when we start at the absorbing state, we may only see a sequence of ones (or a sequence of zeros) which is indistinguishable from a Bernoulli model despite being Markovian.

1.1 Power, convex hulls, and de Finetti

We call an e-value powerful if the corresponding test is powerful, and of course we desire a test that is consistent, meaning that its power goes to one with the sample size. Formally, a level- test has asymptotically power one against $\mathcal{P}$ if

$$\inf_{\mathsf{P} \in \mathcal{P} \setminus \mathcal{Q}} \mathsf{P}(\quad < \infty) = 1.$$

(We explicitly exclude $\mathcal{Q}$ from $\mathcal{P}$ above because, for example, $p_{1j0} = p_{1j1} = p$ recovers an i.i.d. Ber(p) sequence, and so $\mathcal{Q} \subset \mathcal{P}$ as stated. Henceforth, it will always be understood that we desire power against $\mathcal{P} \setminus \mathcal{Q}$.) Similarly, a $\mathcal{Q}$-safe e-value (E_t) is said to be consistent, or power one, if

$$\text{for all} \quad \in (0; 1), \quad \inf_{\mathsf{P} \in \mathcal{P} \setminus \mathcal{Q}} \mathsf{P}(\quad < \infty) = 1; \tag{4a}$$

or equivalently if

$$E_t \to \infty; \ \mathsf{P}\text{-almost surely, for every } \mathsf{P} \in \mathcal{P} \setminus \mathcal{Q}. \tag{4b}$$

Let $\overline{\mathcal{Q}}$ represent the set of all exchangeable distributions over infinite binary sequences. Note that $\mathcal{Q}$ is a rich composite class of parametric distributions, whose convex hull is $\overline{\mathcal{Q}}$ (de Finetti's theorem). Thus, a

sequential test for the i.i.d. setting is also valid under the weaker condition of exchangeability, a fact that we record below, proved in Ramdas et al. [12].

Proposition 1. *The properties of type-1 error control (1) and safety are closed under the convex hull, meaning that any $\mathcal{Q}$-safe e-value is also $\overline{\mathcal{Q}}$-safe, and any level sequential test for $\mathcal{Q}$ is also valid for $\overline{\mathcal{Q}}$.*

As a consequence, we may restrict our attention to developing a $\mathcal{Q}$-safe e-value, and invoke the above fact to step from the i.i.d. setting to the exchangeable setting, and this will be our approach in the rest of this paper. Another consequence, orthogonal to the scope of this paper, is that testing the null $\mathcal{Q}$ against the alternative $\overline{\mathcal{Q}}$ is futile; safe and consistent e-values do not exist and neither do valid, power-one tests.

Remark 2. *To avoid confusion, we note that the convex combination of $\mathbb{Q}, \mathbb{Q}' \in \mathcal{Q}$ must be carefully interpreted. For example, if $\mathbb{Q} = Ber(0.3)^{\infty}$ and $\mathbb{Q}' = Ber(0.7)^{\infty}$ then a draw from $(\mathbb{Q} + \mathbb{Q}')/2$ produces either a sequence with 70% zeros or with 70% ones, each with probability half, and produces a sequence with equal number of zeros and ones with probability zero. Contrast this with the fact that a draw from $(Ber(0.3) + Ber(0.7))/2$ is equally likely to produce a zero or a one. In other words, one must take care to differentiate between $((Ber(0.3) + Ber(0.7))/2)^{\infty}$, which is not in the convex hull of $\mathbb{Q}$ and $\mathbb{Q}'$, and $(\mathbb{Q} + \mathbb{Q}')/2$, which is. Later, we will see that the former lies in the closed* fork-convex hull *of $\mathbb{Q}$ and $\mathbb{Q}'$.*

Note that it is impossible to have a power one test for $\overline{\mathcal{Q}}$ against $\overline{\mathcal{Q}}^c$, since the alternative class is too rich and consists of too many distributions that are too close to $\overline{\mathcal{Q}}$, meaning that there are too many ways to violate exchangeability. For example, it should be apparent to the reader that if the first coin has bias p_1 and every other coin has bias $p \neq p_1$, then the resulting sequence is not exchangeable but we would never be able to reliably detect this deviation. This example relies on ensuring that the information required to detect a deviation from the null is exhausted early on in the sequence. To avoid such pathologies it is necessary to restrict the alternative class in some meaningful way. Markovian alternatives are an attractive choice, balancing the needs of relevant practical motivation, tractable mathematical structure, succinct probabilistic description, and intuitive aesthetic appeal. We focus on the setting of a first-order Markov alternative, and briefly return to address higher order alternatives later.

1.2 Likelihood ratios and martingales

Let the likelihood under a particular $\mathbb{Q} \in \mathcal{Q}$, where $\mathbb{Q} = \text{Ber}(p)^{\infty}$, be represented by

$$Q_t \equiv Q_p(X_1, \ldots, X_t) := (1-p)^{n_0} p^{n_1},$$

where $n_0 = n_0(t)$ and $n_1 = n_1(t)$ represent the number of zeros and ones seen up to time t. The likelihood associated to $\mathbb{P} \in \mathcal{P}$ is given by

$$P_t \equiv P_{p_{1|0}, p_{1|1}}(X_1, \ldots, X_t) := \frac{1}{2} \prod_{s=2}^{t} p_{X_s | X_{s-1}} = \frac{1}{2} p_{0|0}^{n_{0|0}} p_{1|0}^{n_{1|0}} p_{0|1}^{n_{0|1}} p_{1|1}^{n_{1|1}},$$

where $n_{1|0} = n_{1|0}(t)$ is the total number of ones following zeros up to time t, etc.

Naturally, for a point null $\mathbb{Q} \in \mathcal{Q}$ and point alternative $\mathbb{P} \in \mathcal{P}$, *Wald's sequential likelihood ratio test (SLRT)* [21] yields a power-one test. The likelihood ratio process, i.e., (P_t/Q_t), is a $\mathbb{Q}$-martingale starting in one and thus a $\mathbb{Q}$-safe e-value, and the resulting test that thresholds at level $1/\alpha$ has optimal power against $\mathbb{P}$. For composite nulls and alternatives, the SLRT cannot be directly applied. The *mixture SLRT* integrates over the alternatives using a "prior" distribution (or, more appropriately, mixture distribution, to avoid any Bayesian interpretations of our frequentist statements), but this only works for composite alternatives, since mixing over the null set does not yield a safe e-value or the desired type-1 error control property in (1). (We note here that, interestingly, the GROW e-values of [6] are ratios of mixtures, though they are safe only for a fixed sample size.) The generalized SLRT maximizes the likelihood under both null and alternative, but this also does not yield a martingale or a safe e-value. In both cases, it is difficult to find a threshold for the resulting process that achieves type-1 error control in (1), since the SLRT's choice of $1/\alpha$ does not suffice.

Despite the above apparent difficulties in generalizing the SLRT to yield an e-value, it has been recently established that nonnegative (super)martingales play a fundamental role in the design of admissible sequential tests (and the construction of admissible safe e-values), even for composite nulls [12]. In anticipation of the results to follow, it is useful to set up some relevant notation. In what follows, a process $(M_t)_{t\ 0}$ will be called a $\mathbb{Q}$-NM if it is a nonnegative martingale with initial value one, that is, (M_t) is adapted to $(\mathcal{F}_t)$, $M_0 = 1$ and $\mathsf{E}_{\mathbb{Q}}[M_t|\mathcal{F}_s] = M_s \geq 0$ for any $s \leq t$. Such processes are called test martingales by Shafer et al. [15]. If (M_t) is a $\mathbb{Q}$-NM simultaneously for every $\mathbb{Q} \in \mathcal{Q}$, then we will call it a $\mathcal{Q}$-NM. If the equality above is replaced by an inequality $\leq$, then we will call it a $\mathbb{Q}$-NSM or $\mathcal{Q}$-NSM (nonnegative supermartingale). An appropriate variant of the optional stopping theorem implies that for any $\mathcal{Q}$-NSM (M_t) and any stopping time $\quad$ (potentially infinite), the stopped process has expectation at most one, or in other words

$$\sup_{\mathbb{Q}2\mathcal{Q}} \sup_{\quad 2\mathcal{T}} \mathsf{E}_{\mathbb{Q}}[M\ ] \leq 1.$$

Indeed, it is well known that for any $\mathcal{Q}$-NSM, $M_{\ 1} := \lim_{t\ !\ 1} M_t$ is a well defined random variable, and $\sup_{\mathbb{Q}2\mathcal{Q}} \mathsf{E}_{\mathbb{Q}}[M_{\ 1}] \leq 1$. The correspondence with the definition of a safe e-value is not coincidental — to construct a $\mathcal{Q}$-safe e-value, it suffices to construct a $\mathcal{Q}$-NSM. However, we claim the following.

Proposition 3. *Every $\mathcal{Q}$-NSM is also a $\mathcal{P}$-NSM (recall that $\mathcal{Q}$ and $\mathcal{P}$ contain all i.i.d. respectively first-order Markov distributions). In other words, any $\mathcal{Q}$-safe e-value with nontrivial power cannot be a $\mathcal{Q}$-NSM, since the latter is powerless against $\mathcal{P}$ by virtue of being a $\mathcal{P}$-NSM.*

This paper is as much about understanding the above negative result, as about providing a positive result. In other words, this result probes at the "gap" between a $\mathcal{Q}$-safe e-value and a $\mathcal{Q}$-NSM. The former is a much weaker property than the latter. While the latter suffices for the former, it is by no means necessary, as recently observed in a more abstract setup [12]. Indeed, a $\mathcal{Q}$-NSM (N_t) satisfies the much stronger "conditional" property that

$$\mathsf{E}_{\mathbb{Q}}[N\ |\mathcal{F}_s] \leq N\ _{\wedge s} \text{ for every } \mathbb{Q} \in \mathcal{Q} \text{ and stopping time } \quad \in \mathcal{T};$$

which implies the earlier mentioned optional stopping result. Indeed, the above property is satisfied if and only if (N_t) is a $\mathcal{Q}$-NSM, but the earlier properties can be satisfied even by processes that are upper bounded by NSMs, but are not themselves NSMs. It is exactly this gap that we will exploit.

We provide a geometrical characterization of the above phenomenon: essentially, we will show that the above proposition is true because $\mathcal{P}$ lies within the "fork-convex hull" of $\mathcal{Q}$, and we prove that this hull preserves the NSM property of a process. Thus, a $\mathcal{Q}$-NSM yields a powerless test against $\mathcal{P}$ since it is automatically and unintentionally safe under the alternative as well as the null. Along the way, we will encounter other friends from martingale theory, such as the Snell envelope, which previously has not played a prominent role in the mathematical treatment of sequential testing.

1.3 A powerful $\mathcal{Q}$-safe e-value that is not a $\mathcal{Q}$-NSM

Ramdas et al. [12] show that any $\mathcal{Q}$-safe e-value is dominated by a $\mathcal{Q}$-safe e-value of the form

$$E_t := \inf_{\mathbb{Q}2\mathcal{Q}} M_t^{\mathbb{Q}};$$

where $(M_t^{\mathbb{Q}})$ is a $\mathbb{Q}$-NM. As mentioned before, each limiting variable $M_{\ 1}^{\mathbb{Q}}$ is well defined and has expectation at most one, and thus, $E_{\ 1} := \limsup_{t\ !\ 1} E_t$ also has expectation at most one. The safety property immediately holds at infinite times as well, meaning that $\sup_{\mathbb{Q}2\mathcal{Q}} \mathsf{E}_{\mathbb{Q}}[E_{\ 1}] \leq 1$ and $\sup_{\ 2\mathcal{T}} \sup_{\mathbb{Q}2\mathcal{Q}} \mathsf{E}_{\mathbb{Q}}[E\ ] \leq 1$. To avoid too much suspense before we get into these subtle new concepts, we first present our solution immediately, and delay its derivation to the next section. To this end, define

$$R_t\ := \frac{n_{0j0} + 0{:}5 \quad n_{0j1} + 0{:}5 \quad n_{1j0} + 0{:}5 \quad n_{1j1} + 0{:}5}{2\ \ (0{:}5)^4\ \ (n_{0j0} + n_{1j0} + 1)\ \ (n_{0j1} + n_{1j1} + 1)} \cdot \quad \frac{n_1}{t}^{\ n_1}\ \frac{n_0}{t}^{\ n_0}\ ; \tag{5}$$

where $\quad$ denotes the usual gamma function.

Theorem 4. *The process (R_t) is a $\mathcal{Q}$-safe e-value, and thus thresholding it at level $1=$ yields a level- sequential test . Furthermore, this test has power one, i.e., (4a) holds.*

Above, (R_t) is not itself a $\mathcal{Q}$-NSM, but is nevertheless upper bounded by a (different!) Q-NM for every $Q \in \mathcal{Q}$, resulting in it being a $\mathcal{Q}$-safe e-value. This idea is enabled by bringing together the method of mixtures (using Jeffreys' prior) for combining the composite alternative, with the maximum likelihood under the composite null. Beyond showing that it has power one, one can quantify that it has rate-optimal power by utilizing a regret bound from universal coding. The next section essentially proves the above theorem, after which we turn to defining fork-convexity. We end with a discussion about this paper's approach compared to other possible approaches to the problem.

2 Jeffreys' mixture meets maximum likelihood

As briefly mentioned earlier, if the null set was a singleton, say corresponding to $= \text{Ber}(p)$, and the alternative was also a singleton, such as when (X_t) deterministically alternates between 0 and 1, then Wald's sequential likelihood ratio test [21] would immediately yield a solution to the problem at hand. To elaborate, let f_p denote the probability mass function of a $\text{Ber}(p)$ random variable and let L_{01} denote the likelihood function under the alternative:

$$L_{01}(X_1, \dots, X_t) := \begin{cases} 1 & \text{if } (X_1, X_2, \dots) = (0, 1, 0, 1, \dots); \\ 0 & \text{otherwise.} \end{cases}$$

Then, for any point null (indexed by p) define the following likelihood ratio:

$$R_t^p := \frac{L_{01}(X_1, \dots, X_t)}{\prod_{s=1}^t f_p(X_s)}; \text{ which equals } \begin{cases} \frac{1}{p^{\lfloor t/2 \rfloor}(1-p)^{\lceil t/2 \rceil}} \text{ with } {}^1\text{-probability } p^{\lfloor t/2 \rfloor}(1-p)^{\lceil t/2 \rceil}; \\ 0 \text{ with } {}^1\text{-probability } 1 - p^{\lfloor t/2 \rfloor}(1-p)^{\lceil t/2 \rceil}. \end{cases}$$

It is easy to check that (R_t^p) is a $\text{Ber}(p)$-NM, and thus a $\text{Ber}(p)$-safe e-value, and from (2) yields a valid level- sequential test that coincides with Wald's original proposal [21]. The question is how to generalize this approach to deal with a composite null and a composite alternative in a computationally tractable and statistically powerful manner.

The following observation deals the first blow: the only process that is a nonnegative martingale under every i.i.d. Bernoulli sequence is one that is almost surely constant. In other words, the only $\mathcal{Q}$-NM is such that $M_t = 1$ for all $t \in \mathbb{N}_0$. This obviously results in a powerless test. So we then turn our attention to constructing a $\mathcal{Q}$-NSM, or a test supermartingale. Unfortunately this approach is dealt a fatal blow by Proposition 3. As alluded to in the intro, we cannot employ mixtures in both numerator and denominator because it violates safety by lowering the denominator too much, and we cannot maximize the likelihood in the numerator and denominator because it violates safety by raising the numerator too much.

Our proposal combines a suitably chosen mixture in the numerator with maximum likelihood in the denominator, thus avoiding both pitfalls.

2.1 Dealing with the composite null via maximum likelihood

To start, let us return to the point alternative described above (alternating 0 and 1), and just handle the composite null using maximum likelihood estimation, as proposed in universal inference [23]. To elaborate, observe that

$$R_t^{\text{ML}} := \inf_{p \in [0,1]} R_t^p = \frac{\text{likelihood under the point alternative}}{\text{maximum likelihood under the null}}$$

is a $\mathcal{Q}$-safe e-value. Indeed, suppose the data truly comes from $\text{Ber}(p^*)$ for an unknown p^*. Then, it is obvious that $R_t^{\text{ML}} \le R_t^{p^*}$, where the latter process is a $\text{Ber}(p^*)$-NM. Thus, for any $Q \in \mathcal{Q}$ (corresponding to some $p^* \in [0, 1]$) and any stopping time $\in \mathcal{T}$, we have

$$\mathbb{E}_Q[R^{\text{ML}}] \le \mathbb{E}_Q[R^{p^*}] \le 1;$$

where the last step invokes the optional stopping theorem for the Ber(p)-NM (R_t^p).

To see that the resulting test has good power, note that under the alternative, $p = 1/2$ uniquely achieves the above infimum at any even time $t \in 2\mathbb{N}$, in which case the denominator equals $(1/2)^t$ and the numerator equals one. Thus, $R_t^{\mathrm{ML}} = 2^t$ at even times, and the test $\mathbb{1}_{\{\sup_{s \le t} R_s^{\mathrm{ML}} \ge 1/\alpha\}}$ (which equals zero until time $1/\alpha$ in (1) and then equals one) is a valid level-α sequential test that stops either at time $\lceil \log(1/\alpha)/\log(2) \rceil$ or at time $\lceil \log(1/\alpha)/\log(2) \rceil + 1$.

To recap, despite the fact that we cannot find a $\mathcal{Q}$-NM, (R_t^{ML}) is a powerful $\mathcal{Q}$-safe e-value against the considered point alternative. This test takes the ratio of the likelihood under the alternative to the maximum likelihood under the null. Next, we detail how to handle the composite alternative $\mathcal{P}$ when testing a point null in $\mathcal{Q}$.

2.2 Dealing with the composite alternative using a Jeffreys' mixture

Taking independent Jeffreys' priors (with densities $w(p) = 1/(\pi\sqrt{p(1-p)})$) for $p_{1|0}$ and $p_{1|1}$, we obtain the mixture likelihood

$$\begin{aligned} \mathsf{P}_{w\,w}(X_1, \ldots, X_t) &:= \int \mathsf{P}_{p_{1|0}, p_{1|1}}(X_1, \ldots, X_t) w(p_{1|0}) w(p_{1|1}) \, \mathrm{d}(p_{1|0}, p_{1|1}) \\ &= \frac{n_{0|0} + 0.5}{2} \cdot \frac{n_{0|1} + 0.5}{(0.5)^4} \cdot \frac{n_{1|0} + 0.5}{(n_{0|0} + n_{1|0} + 1)} \cdot \frac{n_{1|1} + 0.5}{(n_{0|1} + n_{1|1} + 1)}. \end{aligned}$$

Thus, for any point null represented by Ber(p), we can define the mixture likelihood ratio

$$R_t^{\mathrm{JP}} := \frac{\mathsf{P}_{w\,w}(X_1, \ldots, X_t)}{\prod_{s=1}^t f_p(X_s)} = \frac{\text{Jeffreys' mixture over the alternative}}{\text{likelihood under the point null}}.$$

Using Fubini's theorem to swap integrals, it is easy to check that R_t^{JP} is a Ber(p)-NM, and the corresponding sequential test is Wald's usual mixture SLRT [22]. Note that it is the very particular form of this mixture that yields a closed form expression and thus a computationally feasible test. However, we do not use this mixture just for computational reasons; as we detail soon, combining it with the earlier maximum likelihood idea also yields a statistically near-optimal power.

2.3 Combining Jeffreys' prior with maximum likelihood

Using, as in the previous example, that the likelihood under the null is maximised at $p = n_1/n$, where it evaluates to $(n_1/t)^{n_1} (n_0/t)^{n_0}$, we find that

$$R_t := \frac{\text{Jeffreys' mixture over the alternative}}{\text{maximum likelihood under the null}}$$

reduces to the expression in (5). This is a $\mathcal{Q}$-safe e-value by combining the arguments used for the safety of (R_t^{JP}) and (R_t^{ML}): swapping the maximum likelihood with the (unknown) true likelihood, and then employing Fubini's theorem.

We remark that any prior above would have yielded a $\mathcal{Q}$-safe e-value, and in fact any Beta prior would have yielded one in closed form, but the Jeffreys' prior above allows us to invoke an appropriate optimal regret bound from the universal coding literature [11] to Markov sources (see [17] for a discussion of the resulting optimality):

$$\begin{aligned} R_t &\ge \frac{\frac{1}{2} \left(\frac{n_{1|0}}{n_{1|0}+n_{0|0}}\right)^{n_{1|0}} \left(\frac{n_{0|0}}{n_{1|0}+n_{0|0}}\right)^{n_{0|0}} \left(\frac{n_{1|1}}{n_{1|1}+n_{0|1}}\right)^{n_{1|1}} \left(\frac{n_{0|1}}{n_{1|1}+n_{0|1}}\right)^{n_{0|1}}}{\left(\frac{n_1}{t}\right)^{n_1} \left(\frac{n_0}{t}\right)^{n_0}} \times e^{-\frac{1}{2}\log(n_{1|0}+n_{0|0}) - \frac{1}{2}\log(n_{1|1}+n_{0|1}) - O(1)} \\ &= \frac{\text{maximum likelihood of Markov model}}{\text{maximum likelihood of Bernoulli model}} \times e^{-\log t - O(1)}. \end{aligned}$$

That is, R_t starts gathering evidence against the null if the maximum likelihood for the first-order Markov chain outperforms the maximum likelihood for the Bernoulli model by a factor of order t. Note that

this is a small hurdle to overcome, as the first term is growing exponentially fast in t when the data are explained better by a Markov model, as argued next.

Theorem 5. *Under any first-order non-absorbing Markov alternative whose transition probabilities $p_{1j0}; p_{0j0}; p_{1j1}; p_{0j1}$ satisfy $p_{1j0} \neq p_{1j1}$ (and thus also $p_{0j0} \neq p_{0j1}$) we have $R_t \to \infty$ almost surely.*

The condition on the transition probabilities means that the Markov chain does not reduce to an i.i.d. Bernoulli sequence. Recall that our definition of $\mathcal{P}$ disallows absorbing states. The latter condition is necessary for a power one test, because if (say) 1 is an absorbing state then there is positive probability of seeing only ones (recall that the Markov chain starts at 0 or 1 with equal probability). This is indistinguishable from a realization of an i.i.d. Ber(1) sequence.

Proof of Theorem 5. To simplify notation we define

$$\hat{p}_{1j0} = \frac{n_{1j0}}{n_{1j0} + n_{0j0}}; \qquad \hat{p}_{1j1} = \frac{n_{1j1}}{n_{1j1} + n_{0j1}}; \qquad \hat{p} = \frac{n_1}{t};$$

as well as

$$\hat{q}_{1j0} = \frac{n_{1j0}}{n_1}; \qquad \hat{q}_{1j1} = \frac{n_{1j1}}{n_1}; \qquad \hat{q}_{0j0} = \frac{n_{0j0}}{n_0}; \qquad \hat{q}_{0j1} = \frac{n_{0j1}}{n_0};$$

These quantities all depend on t, although this is suppressed in the notation. Since $p_{1j0} > 0$ and $p_{0j1} > 0$ by assumption, the Markov chain is recurrent and hence the ergodic theorem applies. Then, as t tends to infinity we have $\hat{p}_{1j0} \to p_{1j0}$, $\hat{p}_{1j1} \to p_{1j1}$, and $\hat{p} \to p$, where p is the asymptotic frequency of ones. An expression for p can be obtained by noting that, by definition,

$$p_{1j0} = \frac{\mathsf{P}(X_{t+1} = 1;\ X_t = 0)}{\mathsf{P}(X_t = 0)}; \qquad p_{1j1} = \frac{\mathsf{P}(X_{t+1} = 1;\ X_t = 1)}{\mathsf{P}(X_t = 1)};$$

Hence $p_{1j0}\mathsf{P}(X_t = 0) + p_{1j1}\mathsf{P}(X_t = 1) = \mathsf{P}(X_{t+1} = 1)$. Taking the asymptotic time average of this identity yields the equation $p_{1j0}(1 - p) + p_{1j1}p = p$, which can be re-arranged to

$$p = \frac{p_{1j0}}{p_{1j0} + p_{0j1}}; \tag{6}$$

Next, since $(n_{1j1} + n_{0j1})/n_1 \to 1$ we have

$$\hat{q}_{1j1} = \frac{n_{1j1}}{n_{1j1} + n_{0j1}} \times \frac{n_{1j1} + n_{0j1}}{n_1} = \hat{p}_{1j1} \times \frac{n_{1j1} + n_{0j1}}{n_1} \to p_{1j1};$$

and then, because $\hat{q}_{1j1} + \hat{q}_{1j0} = (n_{1j1} + n_{1j0})/n_1 \to 1$, we get $\hat{q}_{1j0} \to 1 - p_{1j1} = p_{0j1}$. In a similar manner, we get $\hat{q}_{0j0} \to p_{0j0}$ and $\hat{q}_{0j1} \to 1 - p_{0j0} = p_{1j0}$. (The flip from $\hat{q}_{0j1}$ to p_{1j0} is intentional; note also that $n_{0j1} = n_{1j0} \pm 1$, and $n_{1j0} + n_{0j0} \in \{n_0; n_0 + 1\}$, so that $\hat{p}_{1j0} \approx \hat{q}_{0j1}$, and the limiting q matrix is the transpose of the p matrix.)

Let $\ell(t)$ denote the logarithm of the ratio between the maximum likelihood of the Markov model and the maximum likelihood of the Bernoulli model. Using the above notation, this can be written as

$$\ell(t) = n_1 \left[\log \frac{1}{\hat{p}} - \hat{q}_{1j0} \log \frac{1}{\hat{p}_{1j0}} - \hat{q}_{1j1} \log \frac{1}{\hat{p}_{1j1}} \right] + n_0 \left[\log \frac{1}{1-\hat{p}} - \hat{q}_{0j0} \log \frac{1}{1-\hat{p}_{1j0}} - \hat{q}_{0j1} \log \frac{1}{1-\hat{p}_{1j1}} \right];$$

The first parenthesized expression converges, as $t \to \infty$, to

$$\log \frac{1}{p} - p_{0j1} \log \frac{1}{p_{1j0}} - p_{1j1} \log \frac{1}{p_{1j1}};$$

which by Jensen's inequality is greater than or equal to

$$\log \frac{1}{p} - \log \left[p_{0j1} \frac{1}{p_{1j0}} + p_{1j1} \frac{1}{p_{1j1}} \right] = \log \frac{1}{p} - \log \left[\frac{p_{0j1} + p_{1j0}}{p_{1j0}} \right] = 0;$$

using (6) in the last step. Since $p_{1j0} \neq p_{1j1}$ by assumption (otherwise the data would be i.i.d. Bernoulli), Jensen's inequality is actually strict. A similar argument applied to the second parenthesized expression shows that it also converges to a strictly positive number. Therefore, there is a small constant $\varepsilon > 0$ such that for all sufficiently large t we have

$$\ell(t) \geq (n_0 + n_1)\varepsilon = \varepsilon t.$$

Thus, for sufficiently large t, we have $R_t \geq \exp(\varepsilon t - \ln t - O(1)) \to \infty$ almost surely. □

2.4 Extension to higher order Markovian alternatives and context trees

Extensions to Markov sources of order $k > 1$ or alphabet sizes $d > 2$ are immediate. We may treat each k-th order context $x \in \{1, \ldots, d\}^k$ as an independent d-ary prediction problem, and by mixing with independent Jeffreys' (which are Dirichlet$(1/2, \ldots, 1/2)$) priors (or equivalently, composing independent Krichevsky-Trofimov estimators), we obtain a computationally attractive e-value with regret bounded by $(d^k(d-1)/2)\ln t + O(1)$. In other words, we get a closed-form e-value $R_t^{k,d}$ — whose details are tedious, despite being explicit, and thus omitted — such that

$$R_t^{k,d} \geq \frac{\text{maximum likelihood of order } k \text{ Markov model}}{\text{maximum likelihood of Bernoulli model}} \cdot \exp\left(-\frac{d^k(d-1)}{2}\ln t - O(1)\right).$$

The (near)-optimality of this approach is discussed in Takeuchi et al. [17]. The e-value R_t from (5) can be interpreted as $R_t^{1,2}$.

Further computationally attractive extensions include alternatives that consist of Markov sources of varying orders $k = 1, 2, \ldots$ (see discussion on the mixture method for unions below). The even more general Context Tree models have the length of the context that should be taken into account depend on that very context [26].

A similar calculation to the $k = 1, d = 2$ case done previously shows that $R_t^{k,d} \to \infty$, P–almost surely for any alternative $\mathsf{P} \in \mathcal{P}_k \backslash \mathcal{Q}$, where $\mathcal{P}_k$ is the set of Markovian distributions with order at most k. This leads naturally to the following remark.

Remark 6. *Let $\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \ldots$ be a countable sequence of alternatives, that may or may not be nested. Suppose for every $k \in \mathbb{N}$ one can design a safe e-value (E_t^k) for testing $\mathcal{Q}$ against $\mathcal{P}_k$ such that it has power one, meaning that for any $\mathsf{P} \in \mathcal{P}_k \backslash \mathcal{Q}$, we have*

$$E_t^k \to \infty; \ \mathsf{P}\text{–almost surely.}$$

Then, one can design a safe e-value for $\mathcal{Q}$ against $\bigcup_{k \in \mathbb{N}} \mathcal{P}_k$ such that for any $\mathsf{P} \in \bigcup_{k \in \mathbb{N}} \mathcal{P}^k \backslash \mathcal{Q}$, we have

$$E_t \to \infty; \ \mathsf{P}\text{–almost surely.}$$

The proof of the above claim is simple. We can, for example, define the "double mixture"

$$E_t := \sum_{k=1}^{\infty} \frac{6}{\pi^2 k^2} E_t^k,$$

which is a countable mixture over the base e-values (that were already mixed using Jeffreys' prior). It is clear that (E_t) is a safe e-value under $\mathcal{Q}$, by linearity of expectation. To analyze its power, once an alternative P has been picked, let $\mathcal{P}_k$ be the first element of the nested sequence that contains P. Since $E_t^k \to \infty$, P–almost surely, the same property holds for E_t^k / k^2, and thus transfers to E_t since e-values are nonnegative. The computational challenge of calculating E_t remains, but this can be avoided by instead calculating the $\mathcal{Q}$-safe e-value

$$\hat{E}_t := \sum_{k=1}^{t} \frac{6}{\pi^2 k^2} E_t^k.$$

At the (finite) time k, $\hat{E}_t$ begins to include to required term E_t^k, and thus inherits its property of approaching infinity almost surely (consistency). Replacing the sum $\sum_{k=1}^{t}$ by $\sum_{k=1}^{f(t)}$ for any increasing function f that grows to infinity, possibly with sublinear growth (such as $\log(\cdot)$), can further save computation without losing the consistency property.

2.5 Handling generic alternatives by "betting"

An alternative approach to the one above can be found in the recent work on universal inference by Wasserman et al. [23]. It involves a non-anticipating "running" MLE in the numerator, combined with an MLE in the denominator:

$$R_t^{\mathsf{NA}} := \frac{\text{non-anticipating likelihood under the alternative}}{\text{maximum likelihood under the null}}.$$

Here, the numerator (alternative) likelihood is given by

$$\prod_{s=1}^{t} g_s(X_s) \tag{7}$$

where g_s is any "non-anticipating" probability mass function, meaning that it is specified before seeing X_s, but can be learnt using the first $s-1$ data points. Formally, (g_t) must be predictable with respect to $(\mathcal{F}_t)$. One example would be to choose g_s as the (smoothed) maximum likelihood estimator under the alternative using the first $s-1$ samples, but other approaches inspired by machine learning or time series modeling may also be employed.

It is easy to prove that (R_t^{NA}) is a $\mathcal{Q}$-safe e-value: each term can be verified to have conditional mean at most one by swapping the denominator for the (unknown) true null likelihood. The major strength of the above approach is that arbitrarily flexible nonparametric or model-free update rules can be used without sacrificing validity, thus opening up the potential for power against loosely specified alternatives or even the discovery of temporal patterns from the observed data. For example, one may employ a complex Bayesian working model that outputs the posterior predictive probability of observing a zero or one at the next step, and this would not violate any of our theoretical guarantees regardless of the choice of priors or working model. Despite such a strong validity guarantee, the current drawback of this approach is that for generic update rules, there may not be an existing regret bound that we may use to convince ourselves of its power. (Of course, such regret bounds would be available for specific update rules and specific alternatives, and the online learning literature is rapidly expanding the scope and types of available regret bounds for individual sequence prediction.)

As a final remark, this non-anticipating likelihood is closely related to the "predictable-mixture" approach recently explored by [24], and has its roots in Wald [22, Eq. 10:10]. In this vein, it is also closely related to testing hypotheses by betting, as popularized by Shafer and Vovk [13, 14]; specifically (g_t) can be viewed as a sequence of bets on the following outcome.

3 Fork-convexity and $\mathcal{Q}$-Snell envelopes

Forgetting for a moment some of the earlier claims made without proof, one of the main questions we seek to answer in this section is:

> When is a $\mathcal{Q}$-safe e-value simply a $\mathcal{Q}$-NM or $\mathcal{Q}$-NSM in disguise? In other words, is any $\mathcal{Q}$-safe e-value always improved (or recovered) by some $\mathcal{Q}$-NM or $\mathcal{Q}$-NSM?

Such a question was also asked in the latest preprint on safe testing by Grünwald et al. [6]. The necessity and sufficiency results of Ramdas et al. [12] imply that the answer in the singleton $\mathcal{Q} = \{\mathsf{Q}\}$ case is: *always* (via the Doob decomposition of the Snell envelope). The answer in the composite setting is: *sometimes*. We now qualify the 'sometimes' by delving into the rich probabilistic structure underlying safe e-values, examining its relationship to convex null sets, a concept called 'fork-convexity', and a process that we call a 'composite' Snell envelope, known from the mathematical theory of risk measures [3].

Most of this section does not depend on our observations being binary, and we allow the data (X_t) to take values in a more general space $\mathcal{X}$. Some of the technical notions required below, such as local absolute continuity, likelihood ratio (or density) processes, and essential suprema, are reviewed in Appendix A.

3.1 A sequential analog of convexity

We first introduce the concept of *fork-convexity*, which can be viewed as a sequential version of convexity.

Definition 7. *Fix a reference measure* $\mathbb{R}$ *on the sequence space* $\mathcal{X}^{\mathbb{N}}$.

1. *A* fork-convex combination *of two locally dominated laws* $\mathbb{Q}; \mathbb{Q}'$ *with likelihood ratio processes* $(Z_t); (Z'_t)$ *is another law* $\mathbb{Q}''$ *with likelihood ratio process*

$$Z''_t = \begin{cases} Z_t; & t \le s \\ hZ_t + (1-h)Z_s \dfrac{Z'_t}{Z'_s}; & t > s \end{cases} \tag{8}$$

   for some $s \in \mathbb{N}_0$ *and some* $\mathcal{F}_s$*-measurable random variable* h *in* $[0;1]$ *with* $h = 1$ *on* $\{Z'_s = 0\}$. *The latter condition ensures that* (Z''_t) *is well-defined and an* $\mathbb{R}$*-martingale, as required for a likelihood ratio process.*

2. *A set* $\mathcal{Q}$ *of probability measures is called* fork-convex *if every fork-convex combination of elements of* $\mathcal{Q}$ *still belongs to* $\mathcal{Q}$.

Fork-convexity was first introduced by Žitković [20]. It is closely related to a concept in the literature on risk measures called *m-stability*, due to Delbaen [3]. A similar notion called *rectangularity* was introduced by Epstein and Schneider [4] to describe intertemporal preferences with multiple priors. Rectangularity has then been used extensively in the operations research literature in connection with robust Markov decision processes; see e.g. [7, 25, 16].

Note that fork-convexity implies convexity. To see this, observe that any (usual) convex combination $a\mathbb{Q} + (1-a)\mathbb{Q}'$ is also a fork-convex combination; just take $s = 0$ and $h = a$ in (8) to get $Z''_t = aZ_t + (1-a)Z'_t$, which is the likelihood ratio process of $\mathbb{Q}'' = a\mathbb{Q} + (1-a)\mathbb{Q}'$.

A set $\mathcal{Q}_0 = \{\mathbb{Q}\}$ that consists of a single law is clearly fork-convex. A set $\mathcal{Q}_0 = \{\mathbb{Q}^1; \mathbb{Q}^2\}$ consisting of two distinct laws will not be fork-convex; it is not even convex. However, one can form its "fork-convex hull". Here is the general definition.

Definition 8.

1. *The intersection of all fork-convex sets that contain a given set* $\mathcal{Q}_0$ *is called the* fork-convex hull *of* $\mathcal{Q}_0$. *(Note that there is at least one fork-convex set containing* $\mathcal{Q}_0$, *namely the set of* all *laws.)*

2. *The* closed *fork-convex hull of* $\mathcal{Q}$ *is the closure of the fork-convex hull of* $\mathcal{Q}$ *with respect to* $L^1(\mathbb{R})$ *convergence of the likelihood ratio processes at each fixed time* $t \in \mathbb{N}$, *where we recall* $\mathbb{R}$ *is the assumed reference measure.*

Just as for usual convex hulls, the fork-convex hull of $\mathcal{Q}_0$ consists of all finite fork-convex combinations of elements in $\mathcal{Q}_0$. Here a *finite fork-convex combination* of some distributions $\mathbb{Q}^1; \dots; \mathbb{Q}^n \in \mathcal{Q}_0$ is a distribution obtained by iteratively performing (8) a finite number of times on $\mathbb{Q}^1; \dots; \mathbb{Q}^n$, on their fork-convex combinations, on *their* fork-convex combinations, and so on. Closed fork-convex hulls play an important role in Theorem 11 below. Let us illustrate these concepts in a particular example.

Example 9. *Here* $\mathcal{X} = \mathbb{R}$ *and the references measure* $\mathbb{R}$ *is the law under which the data is i.i.d. standard normal (this choice is somewhat arbitrary; we could have chosen any other strictly positive density). Let* $\mathbb{Q}^1; \mathbb{Q}^2$ *be the laws under which* (X_t) *is i.i.d. with* $X_t \sim f_1$ *(under* $\mathbb{Q}^1$*) and* $X_t \sim f_2$ *(under* $\mathbb{Q}^2$*) for some probability density functions* $f_1; f_2$. *The likelihood ratio processes of* $\mathbb{Q}^1$ *and* $\mathbb{Q}^2$ *are*

$$Z^1_t = \prod_{s=1}^{t} \frac{f_1}{\varphi}(X_s); \qquad Z^2_t = \prod_{s=1}^{t} \frac{f_2}{\varphi}(X_s);$$

where φ *is the standard normal density. Given some* $s \in \mathbb{N}_0$ *and* $\mathcal{F}_s$*-measurable random variable* h *in* $[0;1]$ *such that* $h = 1$ *on* $\{Z^2_s = 0\}$, *the corresponding fork-convex combination of* $\mathbb{Q}^1$ *and* $\mathbb{Q}^2$ *is the law* $\mathbb{Q}$

whose density process is

$$Z_t = \prod_{i=1}^{s} \frac{f_1}{r}(X_i) \times \left(h \prod_{i=s+1}^{t} \frac{f_1}{r}(X_i) + (1-h) \prod_{i=s+1}^{t} \frac{f_2}{r}(X_i) \right) :$$

For any Borel set $A \subset \mathbb{R}$ *we have*

$$\mathbb{Q}(X_{s+1} \in A \mid \mathcal{F}_s) = \mathbb{E}_{\mathbb{R}} \left[\frac{Z_{s+1}}{Z_s} 1_A(X_{s+1}) \mid \mathcal{F}_s \right] :$$

A brief calculation using the definition of (Z_t) *as well as the fact that the data is i.i.d. standard normal under* $\mathbb{Q}$ *shows that the right-hand side of the last display evaluates to* $h \int_A f_1(x)\mathrm{d}x + (1-h) \int_A f_2(x)\mathrm{d}x$. *Hence the conditional density of* X_{s+1} *is*

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{Q}(X_{s+1} \le x \mid \mathcal{F}_s) = h(X_1, \ldots, X_s) f_1(x) + (1 - h(X_1, \ldots, X_s)) f_2(x);$$

where we now explicitly indicate that h *depends on* $X_1, \ldots, X_s$. *For* $s = 0$ *this simply means that* X_1 *follows an unconditional mixture,* $\mathrm{d}/(\mathrm{d}x)\, \mathbb{Q}(X_1 \le x) = h f_1(x) + (1-h) f_2(x)$ *for some* $h \in [0, 1]$. *Repeating the above reasoning a finite number of times for different choices of* s *and* h *(and swapping* $\mathbb{Q}^1$ *and* $\mathbb{Q}^2$*) produces the fork-convex hull* $\mathfrak{Q}$ *of* $\{\mathbb{Q}^1, \mathbb{Q}^2\}$. *In summary,* $\mathfrak{Q}$ *is the set of "finite adapted mixtures" of* $\{\mathbb{Q}^1, \mathbb{Q}^2\}$. *More precisely,* $\mathfrak{Q}$ *consists of all probability measures* $\mathbb{Q}$ *such that the conditional density of* X_t *is of the form*

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{Q}(X_t \le x | \mathcal{F}_{t-1}) = h_t(X_1, \ldots, X_{t-1}) f_1(x) + (1 - h_t(X_1, \ldots, X_{t-1})) f_2(x); \quad t \in \mathbb{N}; \tag{9}$$

for some $[0, 1]$*-valued functions* $h_t(x_1, \ldots, x_{t-1})$ *indexed by* $t \in \mathbb{N}$ *such that* $h_t(x_1, \ldots, x_{t-1})$ *equals zero (one) if* $f_1(x_i) = 0$ $(f_2(x_i) = 0)$ *for some* $i = 1, \ldots, t-1$. *Since the fork-convex hull* $\mathfrak{Q}$ *only contains* finite *fork-convex combinations, there is a finite time* T *(depending on the particular element* $\mathbb{Q} \in \mathfrak{Q}$*) beyond which the functions* h_t *will either all be equal to zero, or all equal to one.*

The closed *fork-convex hull, as defined earlier, consists of all* $\mathbb{Q}$ *of the form* (9) *without the restriction that* h_t *eventually equals zero or one.*

To provide some intuition for the definitions as applied to the null $\mathcal{Q}$ considered in this paper, one can imagine a more "algorithmic" process of producing distributions in the closed fork-convex hull. First pick any $p_1 \in [0, 1]$ and observe $X_1 \sim \mathrm{Ber}(p_1)$. Then, after observing X_1, pick any p_2, and observe $X_2 \sim \mathrm{Ber}(p_2)$. Continue this process indefinitely. Then, the (p_i) sequence is predictable and the resulting binary sequence has a law that is contained in the closed fork-convex hull of $\mathcal{Q}$.

It may be instructive to consider another simple example. For a fixed $\quad \in [0, 1]$, define $\mathcal{Q}$ as the set of product distributions $\mathbb{Q}$ over infinite $[0, 1]$-valued sequences such that $\mathbb{E}_{\mathbb{Q}}[X_t | \mathcal{F}_{t-1}] = \mathbb{E}_{\mathbb{Q}}[X_t] = \quad$, and define $\mathfrak{Q}$ as the set of distributions $\mathbb{Q}$ (not necessarily of product form) over infinite $[0, 1]$-valued sequences such that $\mathbb{E}_{\mathbb{Q}}[X_t | \mathcal{F}_{t-1}] = \quad$. Then $\mathcal{Q}$ is not fork-convex if $\quad \in (0, 1)$ but $\mathfrak{Q}$ is, and the latter is the closed fork-convex hull of the former. The problem of sequentially estimating $\quad$ in this setup has been recently studied by Waudby-Smith and Ramdas [24].

3.2 No power against fork-convex hulls

Consider a null set $\mathcal{Q}$ locally dominated by a reference measure $\mathbb{R}$. We now establish the interesting fact that e-values based on $\mathcal{Q}$-NSMs are powerless against any alternative in the closed fork-convex hull of $\mathcal{Q}$. We state this formally in Theorem 11 below, but the underlying reason is contained in the following lemma.

Lemma 10. *If* (L_t) *is a supermartingale under two laws* $\mathbb{Q}, \mathbb{Q}'$*, then* (L_t) *is also a supermartingale under every fork-convex combination* $\mathbb{Q}''$ *of* $\mathbb{Q}$ *and* $\mathbb{Q}'$.

Proof. Note that Q, Q' are dominated by $\mathsf{R} := (\mathsf{Q} + \mathsf{Q}')/2$. Fix any $s \in \mathbb{N}_0$ and $\mathcal{F}_s$-measurable random variable h in $[0,1]$, and let Q'' be the fork-convex combination of Q, Q' given in (8). In compliance with the definition, we restrict h to satisfy $h = 1$ on $\{Z^0_s = 0\}$. Suppose (L_t) is a supermartingale under Q and Q'. Equivalently, $(Z_t L_t)$ and $(Z'_t L_t)$ are supermartingales under R. Thus for $t \in \{1, \dots, s\}$ we have

$$\mathsf{E}_{\mathsf{R}}[Z''_t L_t \mid \mathcal{F}_{t-1}] = \mathsf{E}_{\mathsf{R}}[Z_t L_t \mid \mathcal{F}_{t-1}] \le Z_{t-1} L_{t-1} = Z''_{t-1} L_{t-1}.$$

For $t \ge s+1$ we have

$$\mathsf{E}_{\mathsf{R}}[Z''_t L_t \mid \mathcal{F}_{t-1}] = h\mathsf{E}_{\mathsf{R}}[Z_t L_t \mid \mathcal{F}_{t-1}] + (1-h) Z_s \mathsf{E}_{\mathsf{R}}\left[\frac{Z'_t}{Z'_s} L_t \mid \mathcal{F}_{t-1}\right] \le Z''_{t-1} L_{t-1}.$$

Thus $(Z''_t L_t)$ is an R-supermartingale, or equivalently, (L_t) is a Q''-supermartingale. □

The following theorem refers to the *closed* fork-convex hull of $\mathcal{Q}$. This is the closure of the fork-convex hull of $\mathcal{Q}$, understood in the sense of $L^1(\mathsf{R})$ convergence of the likelihood ratio processes at each fixed time $t \in \mathbb{N}$.

Theorem 11. *Let $\overline{\mathcal{Q}}$ be the closed fork-convex hull of $\mathcal{Q}$. Then every $\mathcal{Q}$-NSM is in fact a $\overline{\mathcal{Q}}$-NSM. Thus a test based on a $\mathcal{Q}$-NSM is powerless against $\overline{\mathcal{Q}} \setminus \mathcal{Q}$.*

Proof. The fork-convex hull of $\mathcal{Q}$ consists of all finite fork-convex combinations of elements of $\mathcal{Q}$. Therefore, thanks to Lemma 10, every $\mathcal{Q}$-NSM remains an NSM under every law Q in the fork-convex hull of $\mathcal{Q}$. To extend this to the closure, pick any element $\mathsf{Q} \in \overline{\mathcal{Q}}$. Then there is a sequence (Q^n) in the fork-convex hull of $\mathcal{Q}$ such that $\mathsf{Q}^n \to \mathsf{Q}$. This means that $Z^n_t \to Z_t$ in $L^1(\mathsf{R})$ for all $t \in \mathbb{N}$, where (Z^n_t) and (Z_t) are the likelihood ratio processes of Q^n and Q, respectively. By passing to a subsequence, we may assume that $Z^n_t \to Z_t$, R-almost surely, for all $t \in \mathbb{N}$. Let (L_t) be any $\mathcal{Q}$-NSM and hence a Q^n-NSM for all n. Equivalently, $(Z^n_t L_t)$ is an R-NSM for all n. By the R-supermartingale property and the conditional version of Fatou's lemma, we get

$$\mathsf{E}_{\mathsf{R}}[Z_t L_t \mid \mathcal{F}_{t-1}] = \mathsf{E}_{\mathsf{R}}\left[\liminf_n Z^n_t L_t \mid \mathcal{F}_{t-1}\right] \le \liminf_n \mathsf{E}_{\mathsf{R}}[Z^n_t L_t \mid \mathcal{F}_{t-1}] \le \liminf_n Z^n_{t-1} L_{t-1} = Z_{t-1} L_{t-1}.$$

This completes the proof that every $\mathcal{Q}$-NSM is in fact a $\overline{\mathcal{Q}}$-NSM. □

The first part of the above theorem asserts that the NSM property is preserved under taking closed fork-convex hulls, but note that this is not true for safe e-values in general. Indeed, (E_t) being $\mathcal{Q}$-safe implies that it is $\mathrm{conv}(\mathcal{Q})$-safe, but not necessarily $\overline{\mathcal{Q}}$-safe.

3.3 Composite Snell envelopes

For a single law $\mathsf{Q} \in \mathcal{Q}$ and an e-value (E_t), the Q-Snell envelope is the smallest Q-NSM that dominates (E_t). It is natural to ask whether, in contrast to this pointwise construction, one can directly construct a "composite $\mathcal{Q}$-Snell envelope", i.e., a smallest $\mathcal{Q}$-NSM that dominates (E_t).

It turns out that the ability to define such a $\mathcal{Q}$-Snell envelope of an e-value depends heavily on the property of fork-convexity. The following result states that if the null set $\mathcal{Q}$ is locally dominated and fork-convex, then a $\mathcal{Q}$-Snell envelope of a given e-value (E_t) exists, is safe, and improves upon (E_t).

Theorem 12. *Let $\mathcal{Q}$ be locally dominated and fork-convex. Let (E_t) be a $\mathcal{Q}$-safe e-value. Then the process*

$$L_t := \operatorname*{ess\,sup}_{\mathsf{Q} \in \mathcal{Q},\ \tau \ge t} \mathsf{E}_{\mathsf{Q}}[E_\tau \mid \mathcal{F}_t], \quad t \in \mathbb{N}_0,$$

where τ ranges over all finite stopping times, is the smallest $\mathcal{Q}$-NSM that dominates (E_t) and satisfies $L_0 \le 1$. Hence, (L_t) is the $\mathcal{Q}$-Snell envelope of (E_t). In particular, by the optional stopping theorem, (L_t) is a $\mathcal{Q}$-safe e-value.

Proof. The proof is essentially a simplified version of an argument due to Delbaen [3, Theorem 11]. This result is argued in continuous time and on a bounded time interval. For the convenience of the reader, we provide a self-contained proof for this paper's discrete-time, infinite-horizon setup. We use properties of the essential supremum reviewed in Appendix A.2, in particular Proposition 16.

For each fixed $s \in \mathbb{N}$, L_s is defined as the essential supremum of the family consisting of all $\mathbb{E}_Q[E\ \ | \mathcal{F}_s]$, indexed by all pairs $(Q;\ )$ with $Q \in \mathcal{Q}$ and $\geq s$ a finite stopping time. We claim that this family is closed under maxima. To prove this claim, let $(Q;\ )$ and $(Q^0;\ ^0)$ be given. Let $A = \{\mathbb{E}_Q[E\ \ | \mathcal{F}_s] \geq \mathbb{E}_{Q^0}[E\ ^0 | \mathcal{F}_s]\}$ and set

$$^{00} = \ 1_A + \ ^0 1_{A^c}$$

and

$$Z_t^{00} = \begin{cases} Z_t; & t \leq s \\ 1_A Z_t + 1_{A^c} Z_s \dfrac{Z_t^0}{Z_s^0}; & t > s \end{cases}$$

where (Z_t) and (Z_t^0) are the likelihood ratio processes of Q and Q^0, respectively. Note that $Z_s^0 > 0$ on A^c so that Z_t^{00} is well-defined. Since A belongs to $\mathcal{F}_s$ and $\ ; \ ^0 \geq s$, ^{00} is a (finite) stopping time. Moreover, since $\mathcal{Q}$ is fork-convex, (Z_t^{00}) is the likelihood ratio process of some $Q^{00} \in \mathcal{Q}$. We now compute

$$\begin{aligned}
\mathbb{E}_{Q^{00}}[E\ ^{00} \mid \mathcal{F}_s] &= \mathbb{E}_{Q^{00}}[1_A E\ \ | \mathcal{F}_s] + \mathbb{E}_{Q^{00}}[1_{A^c} E\ ^0 \mid \mathcal{F}_s] \\
&= \mathbb{E}_R\left[\frac{Z^{00}}{Z_s^{00}} 1_A E\ \ \middle| \mathcal{F}_s\right] + \mathbb{E}_R\left[\frac{Z^{00}\ ^0}{Z_s^{00}} 1_{A^c} E\ ^0 \middle| \mathcal{F}_s\right] \\
&= \mathbb{E}_R\left[\frac{Z}{Z_s} 1_A E\ \ \middle| \mathcal{F}_s\right] + \mathbb{E}_R\left[\frac{Z^0\ ^0}{Z_s^0} 1_{A^c} E\ ^0 \middle| \mathcal{F}_s\right] \\
&= 1_A \mathbb{E}_Q[E\ \ | \mathcal{F}_s] + 1_{A^c} \mathbb{E}_{Q^0}[E\ ^0 \mid \mathcal{F}_s] \\
&= \max\{\mathbb{E}_Q[E\ \ | \mathcal{F}_s]; \mathbb{E}_{Q^0}[E\ ^0 \mid \mathcal{F}_s]\}.
\end{aligned}$$

This demonstrates closure under maxima.

Now fix any $Q \in \mathcal{Q}$ and $s \in \mathbb{N}$. Thanks to the closure property under maxima, Proposition 16 shows that there exist families (Q_n) of measures in $\mathcal{Q}$ and $(\ _n)$ of finite stopping times taking values in $\{s; s+1; \ldots\}$ such that $\mathbb{E}_{Q_n}[E\ _n \mid \mathcal{F}_s] \uparrow L_s$ almost surely under R, and hence under Q. Therefore, by the conditional version of the monotone convergence theorem,

$$\mathbb{E}_Q[L_s \mid \mathcal{F}_{\ 1}] = \mathbb{E}_Q\left[\lim_n \mathbb{E}_{Q_n}[E\ _n \mid \mathcal{F}_s]\ \mathcal{F}_{\ 1}\right] = \lim_n \mathbb{E}_Q[\mathbb{E}_{Q_n}[E\ _n \mid \mathcal{F}_s] \mid \mathcal{F}_{\ 1}]. \tag{10}$$

Replacing Q_n by $(1 - n^{\ 1})Q_n + n^{\ 1}Q$ we still have (10) and, in addition, Q absolutely continuous with respect to Q_n. From now on we use this modified choice of Q_n. Let (Z_t) and (Z_t^n) be the likelihood ratio processes of Q and Q_n, respectively, and define

$$\widetilde{Z}_t^n = \begin{cases} Z_t; & t \leq s; \\ Z_s \dfrac{Z_t^n}{Z_s^n}; & t > s. \end{cases}$$

By fork-convexity, $(\widetilde{Z}_t^n)$ is the likelihood ratio process of some $\widetilde{Q}_n \in \mathcal{Q}$. We then get

$$\begin{aligned}
\mathbb{E}_Q[\mathbb{E}_{Q_n}[E\ _n \mid \mathcal{F}_s] \mid \mathcal{F}_{\ 1}] &= \mathbb{E}_R\left[\frac{Z_s}{Z_{\ 1}} \mathbb{E}_R\left[\frac{Z^n\ _n}{Z_s^n} E\ _n \middle| \mathcal{F}_s\right] \middle| \mathcal{F}_{\ 1}\right] \\
&= \mathbb{E}_R\left[\frac{Z_s}{Z_{\ 1}} \frac{Z^n\ _n}{Z_s^n} E\ _n \middle| \mathcal{F}_{\ 1}\right] \\
&= \mathbb{E}_R\left[\frac{\widetilde{Z}^n\ _n}{\widetilde{Z}_{\ 1}^n} E\ _n \middle| \mathcal{F}_{\ 1}\right] \\
&= \mathbb{E}_{\widetilde{Q}_n}[E\ _n \mid \mathcal{F}_{\ 1}] \\
&\leq L_{\ 1}.
\end{aligned}$$

Combining this with (10) gives $\mathsf{E}_Q[L_s \mid \mathcal{F}_{s-1}] \le L_{s-1}$. Iterating this inequality and using that $\mathcal{F}_0$ is trivial yields $\mathsf{E}_Q[L_s] \le L_0$. In particular, L_s is Q-integrable. Since (E_t) is a $\mathcal{Q}$-safe e-value, we have $L_0 = \sup_{Q \in \mathcal{Q},\, \tau} \mathsf{E}_Q[E_\tau] \le 1$. Since $Q \in \mathcal{Q}$ and $s \in \mathbb{N}$ were arbitrary, this proves that (L_t) is a $\mathcal{Q}$-NSM with $L_0 \le 1$.

Let (L'_t) be another $\mathcal{Q}$-NSM that dominates (E_t). Then for any $Q \in \mathcal{Q}$, any $t \in \{0, 1, \dots\}$, and any finite stopping time $\tau \ge t$, the optional stopping theorem under Q gives $L'_t \ge \mathsf{E}_Q[L'_\tau \mid \mathcal{F}_t] \ge \mathsf{E}_Q[E_\tau \mid \mathcal{F}_t]$. Therefore $L'_t \ge L_t$ by the definition of essential supremum. □

The process (L_t) above is what we call the $\mathcal{Q}$-Snell envelope. Note that the $\mathcal{Q}$-Snell envelope of (L_t) is almost surely equal to (L_t) itself. In short, the above theorem claims that if $\mathcal{Q}$ is fork-convex, then the $\mathcal{Q}$-Snell envelope of any $\mathcal{Q}$-safe e-value exists and is safe.

To construct a powerful and valid test that dominates a safe e-value (E_t), one might be inherently interested in the *largest* $\mathcal{Q}$-NSM $(\overline{L}_t)$ that dominates (E_t) and satisfies $\overline{L}_0 \le 1$. However, we are not aware of a systematic way to obtain such a process. Nevertheless, even the smallest $\mathcal{Q}$-NSM that dominates (E_t), namely the $\mathcal{Q}$-Snell envelope, still tends to improve its power.

For a given $\mathcal{Q}$, can there be more than one process that is considered a $\mathcal{Q}$-Snell envelope (of some other process), and amongst these, is there a largest one? In general, the answer is yes for the first question and (typically) no for the second. Every $\mathcal{Q}$-NSM is its own Snell envelope and there always exist uncountably many $\mathcal{Q}$-NSMs, namely the constant and nonnegative decreasing processes starting at one. In particular, the constant process is also a $\mathcal{Q}$-NM albeit a powerless one. In fact, there may be uncountably many $\mathcal{Q}$-NSMs, with none of these processes dominating the others, and at the same time there may not exist any non-constant $\mathcal{Q}$-NMs (that don't use independent external randomization, which involves expanding the filtration). For this paper's choice of $\mathcal{Q}$, we later show that every $\mathcal{Q}$-NSM is almost surely nonincreasing, and hence the constant process equaling one dominates all $\mathcal{Q}$-NSMs, and indeed the only $\mathcal{Q}$-NM almost surely equals one.

Taken together, Theorems 11 and 12 lead to the following corollary, which tells us that in certain situations one has to move beyond composite NSMs to achieve powerful tests. We continue to let $\mathcal{Q}$ be any locally dominated null set and $\overline{\mathcal{Q}}$ its closed fork-convex hull.

Corollary 13. *Let (E_t) be a $\mathcal{Q}$-safe e-value. Then (E_t) is dominated by (or equals) some $\mathcal{Q}$-NSM (L_t) with $L_0 \le 1$ if and only if (E_t) already happens to be $\overline{\mathcal{Q}}$-safe (and therefore powerless against $\overline{\mathcal{Q}} \setminus \mathcal{Q}$).*

Proof. To prove the forward implication, assume (E_t) is dominated by some $\mathcal{Q}$-NSM (L_t) with $L_0 \le 1$. By Theorem 11, (L_t) is in fact a $\overline{\mathcal{Q}}$-NSM. It follows that (E_t) is $\overline{\mathcal{Q}}$-safe as claimed, because we have $\mathsf{E}_P[E_\tau] \le \mathsf{E}_P[L_\tau] \le L_0 \le 1$ for every $P \in \overline{\mathcal{Q}}$ and each finite stopping time τ.

To prove the reverse implication, assume that (E_t) is actually $\overline{\mathcal{Q}}$-safe. An application of Theorem 12 (with $\mathcal{Q}$ replaced by $\overline{\mathcal{Q}}$) then gives a $\overline{\mathcal{Q}}$-NSM (L_t) with $L_0 \le 1$ that dominates (E_t). This completes the proof of the corollary. □

The above result suggests that we *must* look beyond NSMs for designing sequential tests for exchangeability, and we next show that this fact holds regardless of the class of alternatives considered.

3.4 The inadequacy of NSMs for testing exchangeability

We now return to the main focus of this paper, which is binary sequences; thus $\mathcal{X} = \{0, 1\}$. In this case, any law P is locally dominated by the i.i.d. Bernoulli$(1/2)$ law $R := \mathrm{Ber}(1/2)^\infty$ and the likelihood ratio process of P is

$$Z_t = \prod_{s=1}^{t} 2q_s(X_1, \dots, X_{s-1})\mathbb{1}_{\{X_s = 1\}} + 2(1 - q_s(X_1, \dots, X_{s-1}))\mathbb{1}_{\{X_s = 0\}} \tag{11}$$

for some functions $q_t : \{0, 1\}^{t-1} \to [0, 1]$ such that, R-almost surely,

$$q_t(X_1, \dots, X_{t-1}) = Q(X_t = 1 \mid X_1, \dots, X_{t-1}).$$

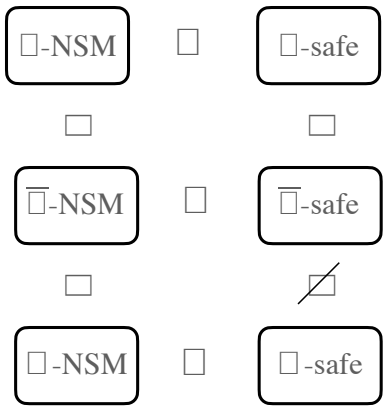

Figure 2: A summary of some of the implications related to $\mathcal{Q}$-NSMs and $\mathcal{Q}$-safety (recall Figure 1 for the definitions of these classes). We would like to design a $\mathcal{Q}$-safe e-value that is powerful against $\mathcal{P}$. Theorem 14 proves that a $\mathcal{Q}$-NSM is non-viable since it unintentionally results in $\mathcal{P}$-safety and thus no power against $\mathcal{P}$. The single non-implication sign above opens a door to constructing a non-NSM based $\mathcal{Q}$-safe e-value that is consistent against $\mathcal{P}$, and this is precisely the construction in (5).

In particular, taking $q_t = p \in (0;1)$ for all t gives the likelihood ratio process of $\mathrm{Ber}(p)^{\infty}$ with respect to R. By repeatedly taking fork-convex combinations of such laws we obtain any law P whose likelihood ratio process is of the form

$$Z_t = \prod_{s=1}^{t} \sum_{k=1}^{T} h_{s\wedge T;k}(X_1,\dots,X_{s-1})\ \big(2p_{s;k}\mathbb{1}_{\{X_s=1\}} + 2(1-p_{s;k})\mathbb{1}_{\{X_s=0\}}\big) \tag{12}$$

for some $N;T \in \mathbb{N}$, some functions $h_{t;k}\colon \{0;1\}^{t-1} \to [0;1]$ with $\sum_{k=1}^{N} h_{s;k} = 1$, and some $p_{t;k} \in (0;1)$. (The T appears because the fork-convex hull only allows for finitely many fork-convex combinations.)

Theorem 14. *Every law P over the space of binary sequences belongs to the closed fork-convex hull of $\mathcal{Q} = \{\mathrm{Ber}(p)^{\infty} : p \in (0;1)\}$. Thus, every $\mathcal{Q}$-NSM must be almost surely nonincreasing, and thus never exceeds one and always has zero power for any .*

Proof. The fork-convex hull of $\mathcal{Q}$ consists of all laws Q obtained by taking fork-convex combinations a finite number of times. In particular, it contains all Q whose likelihood ratio process is of the form (12) with $p_{i;k} \in (0;1)$ and $N;T \in \mathbb{N}$. The *closed* fork-convex hull contains every Q whose likelihood ratio process is of the form (12) for $p_{i;k} \in [0;1]$ and $N = T = \infty$.

Therefore, to prove the theorem it is enough to take an arbitrary law P, whose likelihood ratio process is necessarily of the form (11), and show that it can be written in the form (12). To do so, we must for each $t \in \mathbb{N}$ choose $h_{t;k}$ and $p_{t;k}$ such that

$$q_t(x_1,\dots,x_{t-1}) = \sum_{k=1}^{\infty} h_{t;k}(x_1,\dots,x_{t-1})p_{t;k} \quad \text{for all } (x_1,\dots,x_{t-1}) \in \{0;1\}^{t-1}.$$

This is straightforward: simply let $y_1,\dots,y_{2^{t-1}}$ list all elements of $\{0;1\}^{t-1}$, and set $h_{t;k}(x_1,\dots,x_{t-1}) = \mathbb{1}_{\{f_{y_k}g\}}(x_1,\dots,x_{t-1})$, $p_{t;k} = q_t(y_k)$ for all $k \le 2t-1$ and $h_{t;k}(x_1,\dots,x_{t-1}) = 0$, $p_{t;k} = 0$ for all $k > 2^{t-1}$. □

In other words, not only are $\mathcal{Q}$-NSMs inadequate against Markovian alternatives, they are incapable of detecting *any* deviation from exchangeability.

4 A simulation: power against a change point alternative

We consider a somewhat counterintuitive example to show that a first-order Markov alternative to exchangeability is perhaps more powerful than one may believe at first sight. Consider a length $2n$ sequence of coin flips sampled from $\text{Ber}(p)^n\text{Ber}(q)^n$ for some $p \neq q$. To match the setup of this example with our initial problem set up, one could potentially extend this to an infinite sequence in an arbitrary way, for example just continuing as $\text{Ber}(q)$ after time $2n$.

This sequence is clearly not exchangeable. It is, however, not clear whether our proposed first-order Markov alternative would detect (much) evidence against the null, as the sequence is not Markov, but is more like a change point alternative. Detecting evidence is not a given; the outcomes are in fact independent (albeit not identically distributed). Hence there is no first-order dependency structure for the Markov model to exploit. And on top of that, there seems to be only one problematic time-point, precisely half-way through the sequence. So even if the Markov model somehow exploited this, how could it gain an amount of evidence growing with the length n of the sequence?

We now show that the above arguments are all misguided, and that the process (R_t) from (5) gains an amount of evidence against the exchangeable null that grows exponentially with t, between time n and $2n$. The evolution of (R_t) on a typical run of this process is shown in Figure 3.

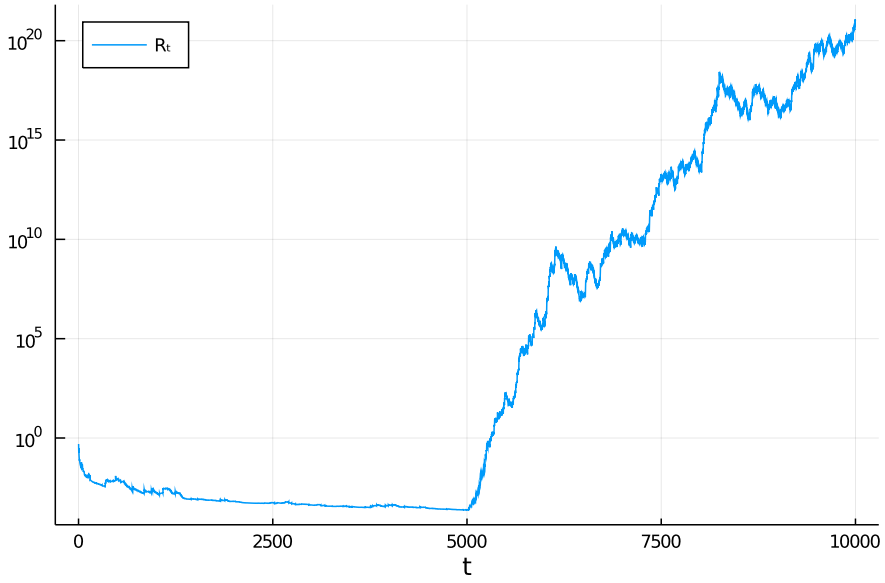

Figure 3: The process (R_t) on a sequence sampled from $\text{Ber}(.1)^{5000}\text{Ber}(.4)^{5000}$. On the first half, we see (R_t) decays as $1/t$, which is due to the overhead of Jeffreys' mixture for the Markov model over the maximum likelihood Bernoulli parameter. After the change point, we see (R_t) increasing fast on the exponential scale. Recalling (3) for those more familiar with the p-value scale, the corresponding anytime p-value dips below 10^{-20} towards the end.

Initially, (R_t) loses steam and tends towards zero at a rate $1/t$ before time n since the null is true and there is a price to pay for the Jeffreys' mixture over the alternative. To calibrate what to expect after the change point, think of n as being relatively large so that we can reason about empirical frequencies of zeros and ones with more ease. Let us compute the maximum likelihood parameters for typical sequences with frequency (tending to) p in the first half and q in the second half. For the Bernoulli model, we find $\hat{p} = (p + q)/2$. For the first-order Markov model we find that

$$\hat{p}_{1|1} = \frac{p^2 + q^2}{p + q} \quad \text{and} \quad \hat{p}_{1|0} = \frac{(1-p)p + (1-q)q}{(1-p) + (1-q)}.$$

The main observation here is that the best Markov model is i.i.d. if $\hat{p}_{1j1} = \hat{p}_{1j0} = \hat{p}$, which occurs if and only if $p = q$. The fact that an exploitable first-order Markov dependency structure arises can perhaps be best observed in the extreme case $p = 0$ and $q = 1$. As this comparison does not really depend on n, we find that for all other parameter settings with $p \neq q$, the Markov model will gain overall evidence exponentially growing with t between time n and $2n$. (Technically, the exponential growth does not start immediately at time $n + 1$, but it does so eventually.) However, as t grows even further — say beyond $t = n^2$ or $t = 2^n$ — R_t will decrease once more towards zero. This is because the sequence eventually is dominated by i.i.d. Ber(q) coin flips, and the MLE under the null explains the data very well.

Thus, for this example, we do not get a power one test, nor should we expect a single change point away from an i.i.d. model to yield power one for a test designed to be powerful against Markovian alternatives. If the initial pattern repeats itself after $2n$ steps, meaning that we keep alternating between Ber(p) and Ber(q) models, then (R_t) does have power one, and this is interesting because (R_t) is designed for first-order Markov alternatives, but it is consistent against these $2n$-th order Markov alternatives.

In fact, one can argue that it is information theoretically impossible to design *any* power one test, including tests that are tuned to detect a single change point. To see why, think of very small n, like $n = 2$, to make the reason intuitively transparent. How can a test possibly have enough evidence with just one or two coin flips before the change point, to know with probability one a change actually did occur? Naturally, the larger the time n of the change point, the higher the power could be of any such test (as it is for our test also), but no test can possibly have power one since there is always some small probability (vanishing with n) that the distribution of the first n coin flips look quite similar to the post-change distribution.

Nevertheless, this simple example illustrates the point that our proposed e-value R_t for evidence against exchangeability is actually powerful even in scenarios that are not (close to) Markov.

4.1 Calibrated p-values and adjusted e-values for not losing capital

While (R_t) is a $\mathcal{Q}$-safe e-value, $(\max_{s \leq t} R_s)$ is not. In other words, we are only allowed to measure our performance based on the wealth accumulated thus far and not the highest wealth that we reached at some point in the process. The same is not true for p-values: $(1/R_t)$ is an anytime p-value, and so is $(1/\max_{s \leq t} R_s)$, the latter being the running infimum of the former. In game-theoretic terminology, the gambler can decide to stop playing the game (betting against the null) according to any stopping rule τ, but once they have stopped, only the final wealth R_τ of the gambler matters, and a nearly bankrupt gambler cannot point to their past wealth as a measure of their proficiency. This subtle point particularly manifests itself in the above example, because with a single change point, (R_t) rises to some amount (above 10^{20} in the figure) and then will shrink back to zero, so if we happen to stop too late, then R_τ could provide only meagre evidence even though it was once astronomically large.

So how can we get around this worrisome issue? We take inspiration from Shafer et al. [15] and use "calibrated p-values" as our e-values. (As a matter of terminology, our use of calibration here can be seen as an $E \to P \to E$ process, but if we skip the middle step entirely, the $E \to E$ direct method has been called "adjustment" by Dawid et al. [2], Koolen and Vovk [10]. We will present it from both angles below to tie some loose ends in the literature together.)

Define $p_t := 1/\max_{s \leq t} R_s$, so that (p_t) is a $\mathcal{Q}$-valid p-value that satisfies (3). Let f be a calibrator [15, 12], which is a nonincreasing function f such that $\int_0^1 f(u)du = 1$. Then $(f(p_t))$ is a $\mathcal{Q}$-safe e-value. It is not hard to check that $f(p_t) \leq R_t$, so there is some price to pay for being able to take the best possible wealth into account. One possible choice for f is given by

$$f(u) := \frac{1 - u + u \ln u}{u(-\ln u)^2} \; ;$$

also see Vovk and Wang [19, Eq. (2)].

In order to do things more directly, let F be an adjuster [15, 2], which is an increasing function F such that $\int_{1/2}^1 F(y)y^{-2}dy = 1$. Then $A_t = F(\max_{s \leq t} R_s)$ yields a $\mathcal{Q}$-safe e-value, and indeed as before $A_t \leq R_t$.

One possible choice for F is given by

$$F(y) := \frac{y^2 \ln 2}{(1+y)(\ln(1+y))^2}.$$

Thus, even if R_t rises sharply and then decreases to zero eventually, $F(R_t)$ does not. In fact, using the F given above in our example with a single change point, and noting that $F(y) \asymp y/(\ln y)^2$ for large y, we see that $A_1 \approx 10^{17}$ even though $R_1 = 0$. Of course, if $R_t \to \infty$ then so does $F(R_t)$, meaning that it does not lose the consistency property against Markovian alternatives.

Thus, at a (squared) logarithmic price to the overall capital, one can be protected against future losses, and for this reason we recommend using $A_t = F(R_t)$ as an e-value if we are uncertain about how close our alternative might be to the idealized Markovian case studied here.

4.2 Deriving other e-values targeted towards detecting change points

For those explicitly interested in powerful tests to detect change point alternatives in the setting of this paper, we briefly describe a powerful test (albeit not a power one test, as already explained above). Essentially, one can combine the ideas in Remark 6 with those in Section 2.5. We let $\mathcal{P}_k$ denote the alternative in which the change point is hypothesized to occur at time $n = 2^k$, though other increasing functions of k may also suffice. We will define an e-value E_t^k for each k and then use a countable mixture over k as the final e-value.

Now, we describe an e-value for a fixed k. Define $(g_s)_{s=1}^n$ to be a smoothed non-anticipating maximum likelihood estimator, calculated using data from time 1 to $s-1$. The smoothing step is simple: add a single fake observation worth half a heads (or half a tails) to the counts when determining the MLE. The smoothing leads to a slight regularization that can be viewed as the maximum-a-posteriori estimate using a Beta$(1/2, 1/2)$ prior, analogous to Krichevsky-Trofimov betting [11]. Similarly, define $(g_s^+)_{s=n+1}^1$ to be the same smoothed non-anticipating maximum likelihood estimator, but calculated using data from time $n+1$ to $s-1$. In both cases, the smoothing also leads to a well-defined function g_1 and g_{n+1}^+, which are effectively treated as a Ber$(1/2)$ model. Finally, define the e-value E_t^k as the ratio of $\prod_{s=1}^{n \wedge t} g_s(X_s) \prod_{s=(n+1)}^{t} g_s^+(X_s)$ to the maximum likelihood under the i.i.d. null. In other words, the denominator is identical to one of R_t, but the numerator has changed because the targeted alternative is now different.

Recalling Section 2.5 (and the final section of Wasserman et al. [23]), it is easy to see that E_t^k is a $\mathcal{Q}$-safe e-value. If a changepoint occurs at time n (and let $k := \lfloor \log_2 n \rfloor$), the e-values E_t^k and E_t^{k+1} will grow exponentially between time n and $2n$. Even with the countable weighting of Remark 6, their exponential growth washes out the inverse polynomial weights, to yield a powerful e-value E_t.

Naturally, many permutations and combinations of these ideas can be used to derive a variety of tests against different kinds of alternatives. We leave further exploration of these variants to future work.

5 Discussion

The celebrated theorem of de Finetti, for which many proofs exist including based on elementary arguments [9], states that all exchangeable binary sequences are mixtures of i.i.d. sequences. In fact, for any exchangeable sequence, the empirical measure $P_t := (1/t) \sum_{s=1}^t X_s$ converges in distribution to a measure supported on $[0, 1]$, and this is the so-called "de Finetti mixing measure" alluded to in the previous sentence. The crux of the matter is that the convex hull of all i.i.d. binary sequences is precisely the set of exchangeable binary sequences. Since the convex hull preserves properties like safety, one can develop tests for the i.i.d. setting and invoke de Finetti to extend the result to the exchangeable setting.

In this paper, we go several steps further: we prove that the set of Markovian sequences lies in the "fork-convex hull" of all exchangeable (or i.i.d.) sequences. In fact, Theorem 14 shows that the closed fork-convex hull is so large that *every* law over binary sequences is contained in it! Theorem 11 shows that the nonnegative supermartingale (NSM) property is preserved not just by taking the convex hull of a set of distributions, but also when taking the (much larger) fork-convex hull, and Corollary 13 shows

that any safe test for $\mathcal{Q}$ is also safe for its fork-convex hull. Together, these results show that any NSM under exchangeable distributions is also an NSM under Markovian distributions, and in fact it is an NSM under *every* distribution over binary sequences, meaning that test statistics that are NSMs are powerless to distinguish non-exchangeable distributions from exchangeable ones.

We get around the above hurdles by designing a process (R_t) in (5) that is upper bounded by some nonnegative martingale for every exchangeable distribution, despite not being an NSM itself. This process uses the method of mixtures with Jeffreys' prior to handle the composite alternative, along with the maximum likelihood under the null, to ultimately yield a computationally efficient closed-form e-value. This e-value not only has the desired safety properties at arbitrary stopping times (potentially infinite), but also has power one against any alternative as implied by a regret bound borrowed from the universal coding literature. Section 2 also presented variations that work for higher order Markovian alternatives, and finally also for even more general, loosely specified alternatives by combining the method of predictable mixtures [24] along with universal inference [23].

An interesting approach towards testing randomness was recently expounded by Vovk [18], which is based on conformal prediction. It replaces the canonical filtration $(\mathcal{F}_t)$ by a poorer filtration $\mathcal{G}_t = (S_1, \ldots, S_t)$ formed by conformal scores, where (informally) the score $S_t \equiv S(X_t; \{X_1, \ldots, X_{t-1}\})$ measures how different (X_t) is from $\{X_1, \ldots, X_{t-1}\}$, in other words how much it does or does not conform to the past. Vovk then produces a sequence of independent p-values under the null, which are converted to e-values by appropriate calibration, which are in turn combined to form a martingale with respect to $(\mathcal{G}_t)$. This is particularly interesting because Vovk argues, much like in our setting, that the only martingales with respect to $(\mathcal{F}_t)$ are almost surely constants, but he is able to identify nontrivial martingales with respect to an appropriately impoverished filtration $(\mathcal{G}_t)$.

Our approaches based on Jeffreys' mixture and the nonanticipating likelihood (or predictable mixture) can be seen as providing two alternatives to Vovk's methodology. In fact, the latter bears some commonalities to Vovk's approach, in that the function $S(\cdot; \{X_1, \ldots, X_{t-1}\})$ mentioned above must be predictable, just like the sequence (g_t) used in (7), which are all connected to betting approaches to statistical inference. Indeed, Vovk's methodology seems most powerful for change point alternatives, making them most similar to the extensions discussed in Section 4. However, in the end, the details appear to be different, and the conceptual principles by which the methods are derived also differ significantly.

A final, alternate approach to this problem could utilize *reverse* martingales and *exchangeable* filtrations. To elaborate, the exchangeable filtration is the reverse filtration $(\mathcal{E}_t)_{t=0}^{1}$ where $\mathcal{E}_0 := \sigma(\{X_1, X_2, \ldots\})$, and for all $t \geq 1$, $\mathcal{E}_t$ denotes the σ-algebra generated by all real-valued Borel-measurable functions $f(X_1, X_2, \ldots)$ which are permutation-symmetric in their first t arguments, so that $\mathcal{E}_0 \supseteq \mathcal{E}_1 \supseteq \mathcal{E}_2 \cdots$. It is known that if the data are exchangeable, then the empirical measure $P_t := (1/t) \sum_{s=1}^{t} \delta_{X_s}$ forms a measure-valued reverse martingale with respect to the exchangeable filtration, in the sense that $(\int g \, dP_t, t \geq 0)_t$, is a reverse martingale for any bounded and Borel-measurable function g [8]. In fact, the converse of this statement also holds true if the sequence (X_t) is stationary [1]. We hope to explore in more detail whether this approach can lead to powerful tests in the future.

Acknowledgments

AR acknowledges NSF DMS grant 1916320.

6 References

[1] Martin Bladt. Characterisation of exchangeable sequences through empirical distributions. *arXiv preprint arXiv:1903.07861*, 2019. 20

[2] A Philip Dawid, Steven de Rooij, Peter Grunwald, Wouter M Koolen, Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Probability-free pricing of adjusted american lookbacks. *arXiv preprint arXiv:1108.4113*, 2011. 18

[3] Freddy Delbaen. The structure of m-stable sets and in particular of the set of risk neutral measures. In *In memoriam Paul-André Meyer: Séminaire de Probabilités XXXIX*, volume 1874 of *Lecture Notes in Math.*, pages 215–258. Springer, Berlin, 2006. 10, 11, 14

[4] Larry G. Epstein and Martin Schneider. Recursive multiple-priors. *J. Econom. Theory*, 113(1):1–31, 2003. ISSN 0022-0531. 11

[5] Hans Föllmer and Alexander Schied. *Stochastic Finance*, volume 27 of *De Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, extended edition, 2004. 23

[6] Peter Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing. *arXiv:1906.07801*, June 2019. 2, 4, 10

[7] Garud N. Iyengar. Robust dynamic programming. *Math. Oper. Res.*, 30(2):257–280, 2005. ISSN 0364-765X. 11

[8] Olav Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Springer Science & Business Media, 2006. 20

[9] Werner Kirsch. An elementary proof of de Finetti's theorem. *arXiv preprint 1809.00882*, 2018. 19

[10] Wouter M Koolen and Vladimir Vovk. Buy low, sell high. *Theoretical Computer Science*, 558:144–158, 2014. 18

[11] Raphail E. Krichevsky and Victor K. Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207, March 1981. 7, 19

[12] Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales, 2020. 2, 4, 5, 10, 18

[13] Glenn Shafer. The language of betting as a strategy for statistical and scientific communication (with discussion). *Journal of the Royal Statistical Society, Series A*, 2020. 10

[14] Glenn Shafer and Vladimir Vovk. *Game-Theoretic Foundations for Probability and Finance*, volume 455. John Wiley & Sons, 2019. 10

[15] Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test Martingales, Bayes Factors and p-Values. *Statistical Science*, 26(1):84–101, February 2011. ISSN 0883-4237, 2168-8745. 5, 18

[16] Alexander Shapiro. Rectangular sets of probability measures. *Operations Research*, 64(2):528–541, 2016. ISSN 0030-364X. 11

[17] Jun'ichi Takeuchi, Tsutomu Kawabata, and Andrew R. Barron. Properties of Jeffreys mixture for Markov sources. *IEEE Trans. Inf. Theory*, 59(1):438–457, 2013. doi: 10.1109/TIT.2012.2219171. URL https://doi.org/10.1109/TIT.2012.2219171. 7, 9

[18] Vladimir Vovk. Testing randomness online. *Statistical Science*, 2021. 20

[19] Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination, and applications. *Forthcoming in the Annals of Statistics*, 2019. 18

[20] Gordan Žitković. A filtered version of the bipolar theorem of Brannath and Schachermayer. *Journal of Theoretical Probability*, 15(1):41–61, 2002. ISSN 0894-9840. 11

[21] Abraham Wald. Sequential Tests of Statistical Hypotheses. *Annals of Mathematical Statistics*, 16(2): 117–186, 1945. 4, 6

[22] Abraham Wald. *Sequential Analysis*. John Wiley & Sons, New York, 1947. 7, 10

[23] Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 2020. ISSN 0027-8424. 6, 10, 19, 20

[24] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *arXiv preprint arXiv:2010.09686*, 2020. 10, 12, 20

[25] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013. ISSN 0364-765X. 11

[26] F. Willems, Y. Shtarkov, and T. Tjalkens. The context-tree weighting method: basic properties. 41: 653–664, 1995. 9

A Additional technical concepts and definitions

A.1 Reference measures and local absolute continuity

Consider a probability space with a filtration $(\mathcal{F}_t)_{t2\mathbb{N}_0}$. Let R be a particular probability measure on $\mathcal{F}_1$; we think of R as a *reference measure*. We now explain the concept of local domination and how it allows us to unambiguously define conditional expectations.

- If P is a probability measure on $\mathcal{F}_1$ and is a stopping time, we write $\mathsf{P}|$ for the restriction of P to $\mathcal{F}$. (This is simply the probability measure on $\mathcal{F}$ defined by $\mathsf{P}|(A) = \mathsf{P}(A)$, $A \in \mathcal{F}$. Think of this as the 'coarsening' of P that only operates on events observable up to time .)
- P is called *locally dominated by* R (or *locally absolutely continuous with respect to* R), if $\mathsf{P}|_t \ll \mathsf{R}|_t$ for all $t \in \mathbb{N}$. We write this $\mathsf{P} \ll_{\mathrm{loc}} \mathsf{R}$. More explicitly, this means that

$$\mathsf{R}(A) = 0 \quad \Rightarrow \quad \mathsf{P}(A) = 0; \qquad \text{for any } A \in \mathcal{F}_t \text{ and } t \in \mathbb{N}:$$

  Local absolute continuity does *not* imply that $\mathsf{P} \ll \mathsf{R}$. However, it does imply that $\mathsf{P}| \ll \mathsf{R}|$ for any finite (but possibly unbounded) stopping time . Indeed, if $A \in \mathcal{F}$ and $\mathsf{R}(A) = 0$, then $A \cap \{ \leq t\} \in \mathcal{F}_t$ for all t, and hence $\mathsf{P}(A) = \lim_{t \uparrow 1} \mathsf{P}(A \cap \{ \leq t\}) = 0$.
- A set $\mathcal{P}$ of probability measures on $\mathcal{F}_1$ is called locally dominated by R if every element of $\mathcal{P}$ is locally dominated by R.
- Any $\mathsf{P} \ll_{\mathrm{loc}} \mathsf{R}$ has an associated *likelihood ratio process* (often also called *density process*), namely the R-martingale (Z_t) given by $Z_t := \mathrm{d}\mathsf{P}|_t / \mathrm{d}\mathsf{R}|_t$. Being a nonnegative martingale, once Z_t reaches zero it stays there. Thus with the convention $0/0 := 1$, the ratios Z / Z_t are well-defined for any $t \in \mathbb{N}$ and any finite stopping time $\geq t$. Note that each Z_t is defined up to R-nullsets, and therefore also up to P-nullsets.
- If $\mathsf{P} \ll_{\mathrm{loc}} \mathsf{R}$ has likelihood ratio process (Z_t), the following 'Bayes formula' holds: for any $t \in \mathbb{N}$, any finite stopping times , and any nonnegative $\mathcal{F}$-measurable random variable Y, one has

$$\mathsf{E}_{\mathsf{P}}[Y \mid \mathcal{F}_t] = \mathsf{E}_{\mathsf{R}} \left[\frac{Z}{Z_t} Y \right] \mathcal{F}_t \; \mathbf{1}_{\{Z_t > 0\}}; \qquad \mathsf{P}\text{-almost surely.}$$

  The right-hand side is uniquely defined R-almost surely (not just P-almost surely), and therefore provides a 'canonical' version of $\mathsf{E}_{\mathsf{P}}[Y \mid \mathcal{F}_t]$. *We always use this version.* This allows us to view such conditional expectations under P as being well-defined up to R-nullsets.

One might ask why we work with *local* domination, rather a 'global' condition like $\mathsf{P} \ll \mathsf{R}$ for all P of interest. The answer is that such a condition would be far too restrictive, as we now illustrate. Let $(X_t)_{t2\mathbb{N}}$ be a sequence of random variables. For each $\in \mathbb{R}$, let P be the distribution such that the X_t become i.i.d. normal with mean and unit variance. By the strong law of large numbers, P assigns

probability one to the event $A := \{\lim_{t \to \infty} \frac{1}{t} \sum_{s=1}^{t} X_s = \}$. Moreover, the events A are mutually disjoint: $A \cap A = \emptyset$ whenever $\neq$. This means by definition that the measures P are all mutually singular. Since there is an uncountable number of them, there cannot exist a measure R such that $P \ll R$ for all . On the other hand, if $P|_t$ denotes the law of the partial sequence $X_1, \ldots, X_t$ for some $t \in \mathbb{N}$, then the measures $P|_t$, $\in R$, are all mutually absolutely continuous. In particular, we could (for instance) use $R = P^0$ as reference measure and obtain $P \ll_{\text{loc}} R$ for all $\in R$.

A.2 Essential supremum

On some probability space, consider a collection $(Y)_{\in A}$ of random variables, where $\mathcal{A}$ is an arbitrary index set. If $\mathcal{A}$ is uncountable, the pointwise supremum $\sup_{\in A} Y$ might not be measurable (not a random variable). Alternatively, it might happen that $Y = 0$ almost surely for every $\in \mathcal{A}$, but $\sup_{\in A} Y = 1$. For this reason, the pointwise supremum is often not useful. Instead, one can use the *essential supremum*.

Proposition 15. *There exists a $[-\infty, \infty]$-valued random variable Y, called the* essential supremum *and denoted by* $\operatorname{ess\,sup}_{\in A} Y$*, such that*

1. *$Y \geq Y$, almost surely, for every $\in \mathcal{A}$,*
2. *if Y' is a random variable that satisfies $Y' \geq Y$, almost surely, for every $\in \mathcal{A}$, then $Y' \geq Y$, almost surely.*

The essential supremum is almost surely unique.

In words, the essential supremum is the smallest almost sure upper bound on (Y). The proposition guarantees that it always exists. In some cases, more can be said: the essential supremum can be obtained as the limit of an increasing sequence.

Proposition 16. *Suppose (Y) is closed under maxima, meaning that for any ; $\in \mathcal{A}$ there is some $\in \mathcal{A}$ such that $Y = \max\{Y, Y\}$. Then there is a sequence (n) such that (Y_n) is an increasing sequence and* $\operatorname{ess\,sup}_{\in A} Y = \lim_n Y_n$.

For more information about the essential supremum (and infimum), as well as proofs of the above results, we refer to Section A.5 in [5].