Submitted to Statistical Science

Game-Theoretic Statistics and Safe Anytime-Valid Inference

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk and Glenn Shafer

Abstract. Safe anytime-valid inference (SAVI) provides measures of statistical evidence and certainty—e-processes for testing and confidence sequences for estimation—that remain valid at all stopping times, accommodating continuous monitoring and analysis of accumulating data and optional stopping or continuation for any reason. These measures are based on test martingales, which are nonnegative martingales starting at one. Since a test martingale is the wealth process of a player in a betting game, SAVI uses game-theoretic intuition, language and mathematics. We report recent advances in testing composite hypotheses and estimating functionals in nonparametric settings, leading to new methods even for nonsequential problems.

Key words and phrases: Test martingales, Ville's inequality, universal inference, reverse information projection, e-process, optional stopping, confidence sequence, nonparametric composite hypothesis testing.

1. INTRODUCTION

Stop when you are ahead. Increase your bet to make up ground when you are behind. This is called martingaling in the casino. It often succeeds in the short or medium term, leading novice gamblers to think they can beat the odds and day traders to think they can beat the market (Dimitrov, Shafer and Zhang, 2022). The same delusion arises in science, where sampling until a significant result is obtained is an important source of irreproducibility.

The fallacy of sampling until a significant result is obtained has been discussed by statisticians at least since the 1940s, when Feller (1940) saw it happening in the study of extra-sensory perception. Anscombe (1954) famously called it "sampling to a foregone conclusion", and this inevitability was also pointed out by Robbins (1952).

But disapproval by statisticians has hardly dented the prevalence of the practice. In one widely publicized example, a team of researchers apparently demonstrated benefits from "power posing" (Carney, Cuddy and Yap, 2010). The lead author later disavowed the conclusion and identified the team's peeking at the data as one of her reasons (Carney, Fact 5):

> We ran subjects in chunks and checked the effect along the way. It was something like 25 subjects run, then 10, then 7, then 5. Back then this did not seem like p-hacking. It seemed like saving money (assuming your effect size was big enough and p-value was the only issue).

Ten years ago, an anonymous survey of over 2000 psychologists found 56% admitting to "deciding whether to collect more data after looking to see whether the results were significant" (John, Loewenstein and Prelec, 2012).

Bayesian inference with a prior defined by a statistician's beliefs before seeing any of the data is not affected by (planned) peeking. Problems quickly arise, however, when default or pragmatic priors are used to test composite null hypotheses. These problems are especially severe for commonly used pragmatic priors that depend on the sample size, covariates, or other aspects of the data (De Heide and Grünwald, 2021).

As emphasized by Johari et al. (2022); Howard et al. (2021); Grünwald, De Heide and Koolen (2023); Shafer (2021); Pace and Salvan (2020), amongst others, we need to go beyond disapproval of peeking, and we instead should give researchers tools to fully accommodate it. The branch of mathematical statistics that enables this, sequential analysis, was brilliantly launched in the 1940s and 1950s by Wald, Anscombe, Robbins, and others. The innovations introduced by Robbins, Darling, Siegmund and Lai included *confidence sequences* that are valid at any and all times and *tests of power one*. But these ideas occupied only a small niche in sequential analysis research until around 2017. Since then, interest has exploded and much conceptual progress has been made in parallel threads, which we attempt to summarize.

Carnegie Mellon University, Pittsburgh, USA (e-mail: aramdas@cmu.edu). Centrum Wiskunde & Informatica, Amsterdam, Netherlands (e-mail: pdg@cwi.nl). Royal Holloway London, UK (e-mail: v.vovk@rhul.ac.uk). Rutgers University, USA (e-mail: gshafer@business.rutgers.edu).

This new methodology differs from traditional statistical testing in the way it quantifies evidence against statistical hypotheses. The traditional approach casts doubt on a hypothesis when a selected test statistic takes too extreme a value. This leads to quantifying evidence against the hypothesis by the *p-value*—the probability the hypothesis assigns to the test statistic being so large. The new methodology instead casts doubt on a hypothesis when a selected nonnegative statistic is large relative to its expected value. Imagining that we bought the statistic for its expected value when we selected it, we call the ratio of its realized to its expected value a *betting score* and take this as a measure of our evidence. In the case of a composite hypotheses, we use the infimum of betting scores for the multiple hypotheses and call this an *e-value*. The sequential analog is an *e-process*—a sequence of e-values that monitor the accumulation of evidence. This permits anytime-valid inference; we can repeatedly decide whether to collect more data based on the current e-value without invalidating later assessments, stopping whenever and for any reason whatsoever. This anytime-validity is a form of *safety*. This safety may come with a price, of course. There may (or may not) be tradeoffs between safety and power; see for example Section 8.2.2.

From a technical point of view, the new methodology is based on the concept of a test martingale, along with its betting interpretation. Although martingales became important in probability theory more than a half-century ago, their potential has still not been fully exploited in statistics, and the new emphasis on *nonnegative* (super)martingales has produced a plethora of powerful new methods. These include confidence sequences for a variety of functionals that can be used with multi-armed bandits and new sequential tests for composite null hypotheses. This responds to the need for rigorous methods in many settings that have emerged with the development of information technology in the past half-century, including "living meta-analysis" and the industrial use of A/B testing and multi-armed bandit experiments.

The new methods can be most clearly presented in the language of game-theoretic probability (Shafer and Vovk, 2001, 2019). Here successive observations are Reality's moves in a game. Two other players move before Reality on each round: Forecaster gives probabilities for the outcome, and Skeptic bets by choosing a real-valued function of the outcome, paying its expected value, and receiving its realized value. If Skeptic always chooses nonnegative functions, then the factor by which he multiplies his money (the ratio of the realized to the expected value) is his "betting score" or "e-value" (Shafer, 2021). If he reinvests his money on each round, the betting scores multiply, producing cumulative betting scores that are products of the betting scores for each round so far. Because Skeptic is a free agent, the option of stopping or continuing or even switching to a different experiment on the next round is intrinsic to the game, and the cumulative betting score or e-value quantifies the evidence against the Forecaster (and his probabilities): Skeptic refutes the odds by making money betting at those odds; more money is more evidence that the odds do not reflect reality.

Betting games often fit statistical practice better than measure-theoretic probability models. In particular, they accommodate fully the opportunistic behavior that we want to allow. George Barnard, in his review of Wald's book on sequential analysis (Barnard, 1947), called for embedding statisticians in the sequential decision-making of experimental scientists, in which each batch of observations is followed by deliberation about whether to stop or to continue, perhaps with a modified experiment. The use of a prespecified stopping time, which prescribes continuing only until a certain data-dependent condition is met, obscures or erases this sequential deliberation, pretending that all the decisions flow from a stopping strategy adopted in advance. Barnard's suggestion is better captured by our game-theoretic framework, where a single stopping rule is replaced by notions of evidence that remain valid at *any* stopping time not specified in advance.

Because most readers will be unfamiliar with game-theoretic probability as developed by Shafer and Vovk (2001, 2019), we use the relatively familiar apparatus of measure theory (filtrations, stopping times, martingales, etc.) and new concepts defined within that apparatus (e-values, e-processes, etc.). Frequently, however, we return to the betting story, where our martingales are wealth processes for Skeptic.

2. CENTRAL CONCEPTS

We begin with a sample space Ω equipped with a filtration $\mathbf{F} \equiv (\mathbf{F}_t)_{t\geq 0}$ (an increasing nested sequence of σ-fields), and a set Π of probability distributions on $(\Omega, \mathbf{F})$. We assume that some distribution $P \in \Pi$ governs our data $X \equiv (X_1, X_2, \dots)$. The variables $X_1, X_2, \dots$ need not be independent and identically distributed (iid) under P. We use X^t as a shorthand for $X_1, \dots, X_t$.

A sequence of random variables $Y \equiv (Y_t)_{t\geq 0}$ is called a *process* if it is adapted to $\mathbf{F}$—i.e., if Y_t is measurable with respect to $\mathbf{F}_t$ for every t. Often $\mathbf{F}_t := \sigma(X^t)$, with $\mathbf{F}_0$ being trivial ($\mathbf{F} = \emptyset, \Omega\}$), and in this case Y_t being measurable with respect to $\mathbf{F}_t$ means that Y_t is a measurable function of $X_1, \dots, X_t$. But $\mathbf{F}$ is sometimes a coarser filtration (we discard information, see e.g. Section 4.1) or a richer one (we add external randomization).[1] Y is called *predictable* if Y_t is measurable with respect to $\mathbf{F}_{t-1}$.

[1] In statistical practice, the filtration is usually coarsened, as explained by Alan Turing (Turing, c 1941, p.1): "When the whole evidence about some event is taken into account it may be extremely difficult to estimate the probability of the event, . . . and it may be better to form an estimate based on a part of the evidence . . ."

A stopping time (or rule) τ is a nonnegative integer valued random variable such that $\{\tau \le t\} \in \mathbf{F}_t$ for each $t \ge 0$. In words: we know at each time whether the rule is telling us to stop or keep going. Denote by $\mathcal{T}$ the set of all stopping times, including ones that may never stop.

When we say we are testing $\mathbf{P}$, we mean that we are testing the null hypothesis H_0 that $P \in \mathbf{P}$. When we say we are testing $\mathbf{P}$ against $\mathbf{Q}$, we mean that the alternative hypothesis H_1 is that $P \in \mathbf{Q}$. Typically, $\mathbf{P}$ and $\mathbf{Q}$ are either non-intersecting or nested subsets of Π.

In the sequel, we leave measurability assumptions and other measure-theoretic details implicit so far as possible.

2.1 E-values

An *e-variable* for $\mathbf{P}$ is a nonnegative random variable E such that $\mathbb{E}_P[E] \le 1$ for all $P \in \mathbf{P}$. Its realized value, after observing the data, is an *e-value*. Often we call E itself an e-value, blurring the distinction between the random variable and its realized value. (The term "p-value" is also used both for random variables and their values.)

When $\mathbb{E}_P[E] = 1$, we call the e-value E a *unit bet against* P. This name evokes a story in which expected values are prices of payoffs; the payoff is $E(X)$ and the price is 1. Mathematically, a unit bet against P is the same thing as a likelihood ratio dQ/dP for some alternative Q. This is elementary when we use probability densities:

- $\mathbb{E}_P[E] = 1$ can be written as $\int E(x)p(x) =1$, so that $q := E \times p$ is a density, and $E = q/p$.
- If q and p are Q's and P's densities, then $\mathbb{E}_P[q/p] = \int p(x)\frac{q(x)}{p(x)}dx = 1$.

Unit bets are simply likelihood ratios when testing a single probability distribution P. The generalization to e-values is needed when dealing with a composite $\mathbf{P}$. We use e-values when data are treated as a batch. Their dynamic counterparts are test martingales and e-processes.

2.2 Test Martingales

A process M is a *martingale* for P if

$$\mathbb{E}_P[M_t \mid \mathbf{F}_{t-1}] = M_{t-1} \tag{1}$$

for all $t \ge 1$. M is a *supermartingale* for P if it satisfies (1) with "=" relaxed to "≤". A (super)martingale is called a *test (super)martingale* if it is nonnegative and $M_0 = 1$.

Game-theoretically, a test martingale for P is the wealth process of a gambler who bets against P. If M is a test martingale for P, then $\mathbb{E}_P[M_t] = 1$ for any $t \ge 0$, and thus each M_t is itself a unit bet against P; it is the factor by which M multiplies its money from time 0 to time t. Similarly, the optional stopping theorem implies that for *any* stopping time τ, even potentially infinite, $\mathbb{E}_P[M_\tau] \le 1$, and thus each M_τ is also an e-value for P.

The correspondence between unit bets against P and likelihood ratios with P as the denominator extends to a related correspondence for test martingales for P. If Q is absolutely continuous with respect to P, we can write

$$\frac{q(X^t)}{p(X^t)} = \frac{q(X_1)}{p(X_1)}\frac{q(X_2|X_1)}{p(X_2|X_1)} \cdots \frac{q(X_t|X^{t-1})}{p(X_t|X^{t-1})}, \tag{2}$$

where $X^t := (X_1, \dots, X_t)$, $p(X^t)$ is P's density for X^t, and $q(X^t)$ is Q's density for X^t. Denote the sequence defined by (2) as M; then M is a test martingale for P, and

$$M_t = \prod_{i=1}^{t} B_i = \frac{q(X^t)}{p(X^t)}, \tag{3}$$

$$\text{where} \quad B_t := \frac{q(X_t|X^{t-1})}{p(X_t|X^{t-1})}. \tag{4}$$

Note that each B_t is a unit bet against P, conditional on $\mathbf{F}_{t-1}$; we call B_t *M's unit bet on round t*.

Test martingales for P are always of the form (3). So choosing a test martingale for P comes down to choosing an alternative Q. In applications, constructing a test martingale for P usually amounts to constructing the numerator $q(X_t|X^{t-1})$ in (4); see Section 3.2. Test supermartingales can also be decomposed in the style of (3), where the B_t are *single-round* e-values (i.e. defined as function on a single outcome X_t) conditional on $\mathbf{F}_{t-1}$.

A process M is a test (super)martingale for $\mathbf{P}$ if it is a test (super)martingale for every $P \in \mathbf{P}$. Such composite test (super)martingales are important in this paper. Composite test martingales decompose as in (3): for every $P \in \mathbf{P}$, there is a Q that is absolutely continuous with respect to P and satisfies $M_t = q(X^t)/p(X^t)$.

Trivially, the constant process $M_t = 1$ is a test martingale for any $\mathbf{P}$, and a decreasing process is a test supermartingale for any $\mathbf{P}$. We call a test (super)martingale *nontrivial* if it is not always a constant (or decreasing) process. There may be no nontrivial test martingales if $\mathbf{P}$ is too large. In this case there may still be nontrivial test supermartingales (Section 5.1), but there may also not be (Section 5.5). For this reason, we also need e-processes.

2.3 E-processes

A family $(M^P)_{P\in\mathbf{P}}$ is a *test martingale family* if M^P is always a test martingale for P. A nonnegative process E is called an *e-process* for $\mathbf{P}$ if there is a test martingale family $(M^P)_{P\in\mathbf{P}}$ such that

$$E_t \le M_t^P \text{ for every } P \in \mathbf{P}, t \ge 0. \tag{5}$$

This type of definition was used by Howard et al. (2020), who used the name "sub-ψ process". In parallel, Grünwald, De Heide and Koolen (2023) implicitly defined an e-process for $\mathbf{P}$, also without using the name "e-process", as a nonnegative process E such that

$$\mathbb{E}[E_\tau] \le 1 \text{ for every } \tau \in \mathcal{T}, P \in \mathbf{P}.$$

In words, E must be an e-value at any stopping time. Ramdas et al. (2020) proved that the two definitions are equivalent and that if $\mathbf{P}$ is “locally dominated”, then *admissible* e-processes (Section 8.2.3) must satisfy

$$E_t = \inf_{P\in\mathbf{P}} M_t^P \tag{6}$$

for some test martingale family $(M^P)_{P\in\mathbf{P}}$. (Technically, the inf above is an “essential infimum”.)

Whereas a test martingale for P is the wealth process of a gambler who bets against P, an e-process for $\mathbf{P}$ reports the minimum wealth across many simultaneous betting games, one against each $P \in \mathbf{P}$, all with the same outcomes $X_1, X_2, \dots$ (Ramdas et al., 2022, Section 5.4).

Gamblers often lose some of their money as they continue to play. Similarly, the evidence against a null hypothesis as measured by a test martingale or e-process may decrease as we collect more data. To obtain a measure of evidence that does not decrease, one can take the running maximum of the e-process and then adjust it back to being an e-process using an adjuster or lookback calibrator; see Shafer et al. (2011); Dawid et al. (2011) and Ramdas et al. (2022, Section 4.7).

2.4 Ville’s Theorem and Ville’s Inequality

The notion of a test martingale was first formulated by Ville (1939), though he simply called it a martingale.

Ville gave a proof, valid for any discrete-time stochastic process P, that an event A has measure zero under P if and only if there is a betting strategy that bets against P and becomes infinitely wealthy if A happens—i.e., a test martingale for P that grows to infinity on all of A. Moreover, $P(A) < \epsilon$ if and only if there is a test martingale for P that exceeds $1/\epsilon$ on all of A.[2] These results have been called *Ville’s theorem* (Shafer and Vovk, 2019, Section 9.1). See Ruf et al. (2022) for a generalization to composite $\mathbf{P}$.

Ville also showed that if M is a test martingale for P, then for any $\lambda \ge 1$,

$$P\left(\sup_t M_t \ge \lambda\right) \le \frac{1}{\lambda}. \tag{7}$$

Ville called this the theorem of gamblers’ ruin; a gambler who begins with unit capital and keeps betting until he wins the casino’s entire capital λ has little chance of succeeding. More recently, the theorem has become known as *Ville’s inequality*. For a self-contained proof see Howard et al. (2020) or Crane and Shafer (2020).

Ville’s inequality extends to statements about composite $\mathbf{P}$: if E is an e-process for $\mathbf{P}$, then for every $\alpha \in (0, 1)$,

$$\sup_{P\in\mathbf{P}} P(\exists t \ge 1 : E_t \ge 1/\alpha) \le \alpha. \tag{8}$$

Equivalently, by Howard et al. (2021, Lemma 3),

$$P(E_\tau \ge 1/\alpha) \le \alpha \text{ for every } \tau \in \mathcal{T}, P \in \mathbf{P}. \tag{9}$$

Ville’s inequality plays a central role in converting e-processes into sequential tests or confidence sequences.

2.5 Sequential Tests and their Families

We consider test martingales and e-processes bona fide measures of evidence, with no need for thresholding. But we may want to make a binary decision based on this evidence. We define a (one-sided) sequential test in terms of rejection decisions like $(0, 0, 0, 0, 1, 1, 1, 1, \dots)$, where a 0 means that there is not yet enough evidence to reject the null, and a 1 means that there is. In this formalization, a level-α sequential test $\psi \equiv (\psi_t)_{t\ge1}$ is an increasing process consisting of 0-1 random variables such that

$$P(\exists t \ge 0 : \psi_t = 1) \le \alpha \text{ for all } P \in \mathbf{P}.$$

Howard et al. (2021, Lemma 3) proved that an *equivalent* definition, with optional stopping made more explicit, is

$$P(\psi_\tau = 1) \le \alpha \text{ for any } \tau \in \mathcal{T}, P \in \mathbf{P}.$$

Above, as in the “power-one tests” of Darling and Robbins (1968), we just keep going if we never reject $\mathbf{P}$. This contrasts with Wald’s original picture, where we may finally decide to accept the null.

It is easy to obtain a sequential test from a test martingale or e-process: simply reject the null (and stop) the first time the process reaches or exceeds $1/\alpha$. Indeed, Ville’s inequality implies that $\psi_t := \mathbf{1}(\sup_{s\le t} M_s \ge 1/\alpha)$ is a sequential test. We call a family $(\psi^P)_{P\in\mathbf{P}}$, where ψ^P is a sequential test for P, a *sequential test family*.

2.6 Anytime-valid p-values

A random variable p is a p-value for $\mathbf{P}$ if $P(p \le u) \le u$ for all $P \in \mathbf{P}$ and $u \in [0, 1]$. Like the e-value, this is a static concept. Anytime-valid p-values are the dynamic counterparts of p-values.

An *anytime-valid p-value* (Johari et al., 2022; Howard et al., 2021) for $\mathbf{P}$ is a process $p := (p_t)_{t\ge1}$ such that $P(p_\tau \le u) \le u$ for any $\tau \in \mathcal{T}, P \in \mathbf{P}, u \in [0, 1]$. Equivalently, $P(\inf_t p_t \le u) = P(\exists t \ge 1 : p_t \le u) \le u$. In other words, with probability at least $1 - u$ an anytime-valid p-value will never drop below u. So decisions to stop an experiment or to continue to collect data based on the current value of an anytime-valid p-value are *safe*; they will not violate type-I error control.

It is easy to check that if M is an e-process for $\mathbf{P}$ then $1/(\max_{s\le t} M_s)$ is an anytime-valid p-value for $\mathbf{P}$. In our framework, test martingales and e-processes are central objects for testing, and sequential tests and anytime-valid p-values take a secondary and derivative role.

2.7 Confidence Sequences

When estimating some property of a distribution, like a mean (or a median), we think of it as a functional $\phi : \Pi \to \Theta$ for some space Θ, which is often a subset of $\mathbb{R}^d$.

A $(1 - \alpha)$-*confidence sequence* (CS) is a sequence $(C_t)_{t\ge0}$ of sets $C_t \subseteq \Theta$ such that

$$P(\forall t \ge 1 : \phi(P) \in C_t) \ge 1 - \alpha \text{ for all } P \in \Pi.$$

[2] Shafer and Vovk (2019) turn this around into a *definition* of probability in the betting game.

As before, Howard et al. (2021, Lemma 3) implies that a mathematically equivalent definition is to require

$$P(\phi(P) \in C_\tau) \geq 1 - \alpha \text{ for all } \tau \in \mathcal{T}, P \in \Pi.$$

This dynamic concept can be contrasted with the concept of a confidence set (or interval). A $(1-\alpha)$ confidence set, as usually defined, is required only to contain $\phi(P)$ with probability $1-\alpha$ for a sample of a fixed size or at a single fixed stopping time rather than at all stopping times. Confidence sequences remain valid under continuous monitoring (or peeking) and optional stopping, but confidence sets require the sample size or the stopping time to be fixed in advance of seeing any data.

One can construct a confidence sequence by inverting a family of sequential tests, or thresholding a test martingale family $(M^P)_{P\in\Pi}$: $C_t := \{\phi(P) : P \in \Pi, M_t^P < 1/\alpha\}$. Sometimes it is easier to construct a test martingale family $(M^\theta)_{\theta\in\Theta}$, where M^θ is a test martingale for $\{P : \phi(P) = \theta\}$. In that case, we would define

$$C_t := \{\theta \in \Theta : M_t^\theta < 1/\alpha\}. \tag{10}$$

2.8 Averaging e-values

We can average e-values. By the linearity of expectations, if E_1 and E_2 are e-values for $\mathbf{P}$, then $(E_1 + E_2)/2$ is as well, even if E_1 and E_2 are dependent (for example, calculated in different ways using the same data). This observation generalizes to any number of e-values, and holds for convex combinations or *mixtures* that are not equally weighted. Of course, the e-values to mix and the weights for the mixture must be chosen without looking at the realized e-values; otherwise we are martingaling. Recently, Wasserman, Ramdas and Balakrishnan (2020) used it to derandomize universal inference. Vovk and Wang (2021) have shown that averaging is an admissible way of combining e-values (for a particular definition of admissibility) without further information about the e-values or their dependence structure. (And it is the only admissible symmetric method if we ignore the possibility of further mixing with the constant e-value 1.)

Test (super)martingales and e-processes can also be mixed, yielding mixture (super)martingales or e-processes. This *method of mixtures* goes back to Ville (1939), Wald (1945), and Robbins (Darling and Robbins, 1968; Robbins and Siegmund, 1974), who employed it in various contexts; see Howard et al. (2021) for recent advances.

2.9 Multiplying e-values

Independent e-values can be combined by multiplication. If $B_1, \ldots, B_n$ are independent e-values for $\mathbf{P}$, then the product $B_1 \cdots B_n$ is also an e-value for $\mathbf{P}$. As we saw in Section 2.2, a product of dependent e-values can also be an e-value. If, for all k, B_k is an e-value for $\mathbf{P}$ conditional on the values of $B_1, \ldots, B_{k-1}$, i.e. if

$$\mathbb{E}[B_k \mid B_1, \ldots, B_{k-1}] \leq 1$$

for $k \geq 1$, then $M_n = \prod_{k=1}^n B_k$ is an e-value for $\mathbf{P}$. The sequence $(M_n)_{n\geq 0}$ is a supermartingale with respect to the filtration generated by the B_k. In fact, in Section 2.2 we encountered, and in Sections 4 and 5 we again encounter, at each time t a random variable S_t which is a single-round e-variable conditional on the past, i.e. for all $P \in \mathbf{P}$

$$\mathbb{E}_P[S_t \mid \mathbf{F}_{t-1}] \leq 1. \tag{11}$$

If (11) holds for all t, then $M_n = \prod_{t=1}^n S_t$ is an e-value for $\mathbf{P}$. The sequence $(M_n)_{n\geq 0}$ is a supermartingale with respect to the (often richer) filtration $\mathbf{F}$.

3. GENERAL PRINCIPLES AND METHODOLOGY

As mathematical statisticians learned nearly a century ago from Jerzy Neyman and E. S. Pearson, the choice of a test of a null hypothesis should be guided by the alternative hypotheses that are considered plausible. How should this work when we are using a test martingale, or more generally a test supermartingale or an e-process?

Intuitively, a good supermartingale should grow (get large) fast under the alternative so that we quickly build up evidence against the null as the sample size increases. So we want a test martingale or e-process with maximal expected rate of growth under the alternative.

In this section, we first focus on testing a simple null $\mathbf{P} = \{P\}$ against a simple alternative $\mathbf{Q} = \{Q\}$ and use this case to develop our understanding of expected rate of growth (Section 3.1). We then move to testing a simple null against a composite alternative (Section 3.2), and to the most difficult case, where even the null is composite (Section 3.3), introducing general methods for constructing e-processes—some based directly on growth rate optimality, some more indirectly.

One danger we want to avoid throughout is an e-process becoming zero. Once this happens, the e-process can never become positive again, and thus it can never recognize later evidence against the null, no matter how strong. This can happen with positive probability under a particular alternative Q only if the e-process's strategy for betting for Q (i.e, the test martingale for P designed to become large if Q is correct; remember that the e-process is an infimum for such test martingales for the different alternatives) is sometimes allowed to bet all its money, thus risking bankruptcy. We call this *betting the farm*, and we insist on choosing e-processes that avoid it.

3.1 Simple Null and Simple Alternative

This is the case where we are testing a probability distribution P against an alternative probability distribution Q. As we saw in Section 2.1, the likelihood ratio dQ/dP is the natural test martingale in this case. (This assumes that Q is absolutely continuous with respect to P.)

What are the advantages of using this natural test martingale? The most important advantage, perhaps, is that it has the greatest expected growth as measured using the *expected logarithmic return*, a measure popularized by Kelly (1956). The name *logarithmic return* is standard in finance and hence appropriate when we consider wealth processes. When E is a unit bet against P, E's logarithmic return is simply $\log E$. The fact that the expected logarithmic return $\mathbb{E}_Q(\log E)$ is maximized by $E := dQ/dP$ can be obtained directly from Gibbs' inequality (Shafer, 2021, p. 413). Because M_t is a unit bet against P whenever M is a test martingale for P, it follows that the cumulative likelihood ratio $E_t := (dP/dQ)(X^t)$ maximizes

$$\mathbb{E}_Q(\log E_t) \tag{12}$$

and hence the growth rate $\mathbb{E}_Q(\log E_t)/t$ for each t. By the same argument, $(dP/dQ)(X^\tau)$ maximizes $\mathbb{E}_Q(\log E_\tau)$ for every stopping time τ. Grünwald, De Heide and Koolen (2023) called the requirement that $\mathbb{E}_Q(\log E_\tau)$ be maximized the *GRO* criterion, for "growth-rate optimality" relative to τ. So we may summarize by saying that *in a simple vs. simple test, the likelihood ratio is GRO*.

Why use the logarithm of E rather than some other increasing function of E? In finance, we average the logarithmic returns for successive time periods rather than the simple percentage returns in order to account for compounding. Kelly (1956) pointed out that this compounding means that logarithmic returns add, and hence the law of large numbers applies, allowing us to gain some foresight about the medium to long run. Breiman (1961) showed that the logarithm has a number of other strong optimality properties, especially in iid settings where the wealth can be made to grow exponentially under the alternative, and this criterion maximizes the exponent. Using Wald's identity as Breiman used it, one can show that, in iid settings, the logarithm asymptotically (as $\alpha \to 0$) minimizes expected time before E reaches a desired threshold such as $1/\alpha$, independently of α. It is also true that we will not "bet the farm" when we choose E to maximize the expected logarithm, whereas this can happen if we maximize the expectation of E itself or some polynomial function of E. See also Shafer (2021), who compares expected logarithmic return to power in the Neyman-Pearson theory: both can be used to ask whether the alternatives for which a test is effective are plausible.

3.2 Simple Null and Composite Alternative

How do we find a good test martingale for P when the alternative $\mathbf{Q}$ is composite? In general, we cannot maximize the expected growth rate under all the distributions in $\mathbf{Q}$. But we can look for an alternative Q such that the test martingale defined by (3) has a reasonably high expected growth rate under any distribution in $\mathbf{Q}$ that fits the data $X_1, X_2, \dots$ reasonably well. Because $X_1, X_2, \dots$ are revealed to us progressively, the natural procedure is to construct this Q progressively. On betting round t, we use the data so far, x^{t-1}, to choose the numerator $q(X_t|X^{t-1})$ in (4). Another way to view this is to imagine the data being drawn from (or best explained by) some unknown $Q^* \in \mathbf{Q}$ and — since we do not know Q^* — to attempt to learn it from the data, at each round t plugging in a $q(X_t|X^{t-1})$ that is an estimate of q^* based on past data X^{t-1}.

3.2.1 The Plug-in Method. This is natural when $\mathbf{Q}$ is a parametric model. We use x^{t-1} to estimate the parameters, and this gives us an estimate $\hat{Q}_t$ of the best fitting (or the 'true') Q^*. So our choice for $q(X_t|X^{t-1})$ is $\hat{q}_t(X_t|X^{t-1})$, where $\hat{q}_t$ is $\hat{Q}_t$'s density. Wald proposed, in passing and without any further analysis, this plug-in method (Wald, 1947, Eq. 10.10); it was subsequently analyzed by Robbins and Siegmund (1974), who connect it directly to the mixture method (introduced in the next subsection). Similar ideas were proposed independently by Dawid (1984) for prequential model validation and by Rissanen (1984) as a predictive version of Minimum Description Length learning. Recently, the plug-in method has been employed by Wasserman, Ramdas and Balakrishnan (2020) in parametric models and Waudby-Smith and Ramdas (2020, 2023) in nonparametric models.

We obtain a test martingale M no matter how we estimate the Q_t. But we should not use maximum likelihood, at least when data are discrete, lest we end up betting the farm (the maximum likelihood estimator may assign probability 0 to an outcome that may very well occur in the next round. Most of the authors just cited have found, however, that it often suffices to slightly smooth the maximum likelihood estimator (often using a "prior") to avoid this problem, even in nonparametric settings.

3.2.2 The Mixture Method. Another way to choose the $q(X_t|X^{t-1})$ in (4) is to average over the corresponding conditional distributions for the distributions in $\mathbf{Q}$. We can vary the weights with t and with X^{t-1}. This is the mixture method. The mixture method is not a special case of the plug-in method, because the mixed probability distribution Q we obtain may not be in $\mathbf{Q}$ (this may happen if $\mathbf{Q}$ is not "fork-convex", to use a concept introduced in Ramdas et al. 2022). Since most models used in mathematical statistics are not fork-convex, Q is rarely in $\mathbf{Q}$.

3.2.3 Bayes Factors. We can use a probability distribution on $\mathbf{Q}$, say R, to define weights for a mixture martingale. We update R on each round in the usual Bayesian way. On round t, we use the update $R(\cdot|x^{t-1})$ in the averaging that produces $q(X_t|X^{t-1})$.

Not surprisingly, a simple calculation shows that the resulting unit bet $M_t = \prod_{i=1}^t B_i$ is equal to the Bayes factor defined by the distribution R on $\mathbf{Q}$ (Grünwald, De Heide and Koolen, 2023). This Bayes factor is the ratio $q(X^t)/p(X^t)$, where q is the density from mixing the distributions in $\mathbf{Q}$ with R. But for each t, the conditional probabilities given X^{t-1} obtained by mixing with R are the same as the conditional probabilities given X^{t-1} obtained with $R(\cdot|X^{t-1})$. Conditioning on the data so far commutes with averaging the distributions in the model.

Bayes factors have been advocated by many statisticians as measures of evidence against a null when the alternative is composite (Berger, Pericchi and Varshavsky, 1998; Jeffreys, 1961). E-values measure evidence against the null in a different way. Whereas a Bayes factor is used to multiply prior odds, an e-value is intuitively the outcome of a bet. Not surprisingly then, the correspondence between mixture test martingales (or e-values) and Bayes factors does not extend to composite nulls.

3.2.4 Minimizing the Worst. If we do not have a priori knowledge to guide us when determining the 'prior' distribution R in the method of mixtures, we may look for the distribution R that minimizes the worst possible shortfall from this best growth rate. This means that we measure the quality of test martingale M stopped at time τ by

$$\inf_{Q \in \mathbf{Q}} \mathbb{E}_Q(\log M_\tau - \log M_\tau^Q). \tag{13}$$

We want this nonpositive quantity to be as large (as close to zero) as possible. Grünwald, De Heide and Koolen (2023) introduced this criterion and called it REGROW (*RElative GRowth Optimality in Worst case*). Under slight regularity conditions, the e-variable M_τ that maximizes (13) can be written as a Bayes factor defined relative to a specific prior R_τ, where R_τ varies with τ. Still, in the cases considered by Grünwald, De Heide and Koolen (2023) one can find priors R that get us close to the maximum for all τ. In some settings we do find a unique test martingale that maximizes (13) for all τ. We will generalize (13) to the case of composite nulls below, and we will find a unique e-process that maximizes the criterion for all τ for the group invariance tests of Section 4.1.

Grünwald, De Heide and Koolen (2023) show that when $\mathbf{Q}$ is an i.i.d. exponential family and $\mathbf{P} = \{P\}$ is simple, the test martingale M obtained from Jeffreys' prior is asymptotically REGROW: for every τ set equal to a large t, M_t maximizes (13) up to an $o(1)$ term among all e-variables that can be defined on X^t, linking growth optimality to description length (Section 8.1.3).

3.2.5 The (Smoothed) Empirical Distribution as Alternative. Waudby-Smith and Ramdas (2023) have proposed a related method, which they call GRAPA (Growth Rate Adaptive to the Particular Alternative). On round t, it uses the empirical distribution of X^{t-1}, possibly smoothed, for the numerator in (4). In practice, this yields efficient tests and (by inversion) confidence sets. GRAPA tries to mimic the growth rate that would be achieved by using a smoothed empirical distribution as the alternative (or its projection onto $\mathbf{Q}$ when possible), while REGROW tries to match the Q^*-expected growth rate by learning Q^* by the method of mixtures.

3.3 Composite Null and Alternative

When the null $\mathbf{P}$ and alternative $\mathbf{Q}$ are both composite, we can usually handle them in a modular fashion. The composite alternative can be handled as in the previous section, using the plug-in method or the method of mixtures. Here we describe two relatively general ways of handling the composite null: universal inference (UI)—which always yields an e-process—and reverse information projection (RIPr)—which yields a sequence of e-variables that is sometimes an e-process. As mentioned in Section 2.2, test (super)martingales for composite $\mathbf{P}$ may not exist. So we use the general concept of an e-process.

UI and RIPr are not the only ways of handling composite nulls. In Section 5, we will see many other test (super)martingales and e-processes for composite nulls. Most of these involve the method of mixtures applied directly to a collection of e-variables rather than to distributions in $\mathbf{Q}$, as briefly introduced in Section 3.3.3 below.

3.3.1 Universal Inference (UI). This method, introduced by Wasserman, Ramdas and Balakrishnan (2020) uses e-processes of the form

$$E_t^{\mathrm{UI}} := \frac{\bar{q}(X^t)}{\sup_{p \in \mathbf{P}} p(X^t)} = \frac{\bar{q}(X^t)}{\hat{p}_{X^t}(X^t)}, \tag{14}$$

where $\bar{q}(X^t) := \prod_{i=1}^t \hat{q}_{X^{i-1}}(X_i)$, $\hat{q}_{X^{i-1}}$ is any distribution learnt from X^{i-1}, and $\hat{p}_{X^t}$ is the maximum likelihood estimator (MLE) under $\mathbf{P}$, the final equality holding whenever the MLE is well-defined. Alternatively, we can use the method of mixtures and set $\bar{q}(X^t) := \int \prod_{i=1}^t q(X_i \mid X^{i-1}) dR(q)$, where R is a distribution over $\mathbf{Q}$. In either case, as in the preceding subsection, the numerator is equal to $\bar{q}(x^t)$ for some alternative $\bar{Q}$ (usually not in $\mathbf{Q}$), so that E_t^{UI} is the infimum of the family of test martingales $(\bar{q}(X^t)/p(X^t))_{p \in \mathbf{P}}$ and hence an e-process by (5). The method is *universal* because it does not require regularity assumptions or asymptotics and, importantly, is applicable in both parametric and nonparametric settings; see Section 5.5 for an example of each.

We can think of E_t^{UI} as a middle ground between the non-Bayesian generalized likelihood ratio (MLE in both numerator and denominator) and the Bayes factor for a composite null (mixtures in both numerator and denominator), neither of which leads to an e-process in general. By taking a supremum in the numerator, the generalized likelihood ratio exaggerates evidence for the alternative, requiring that this exaggeration be taken into account using the ratio's sampling distribution. By including poorly fitting distributions in its mixture in the denominator, the Bayes factor may downplay evidence for the null.

3.3.2 Reverse Information Projection (RIPr). This method, pioneered by Grünwald, De Heide and Koolen (2023), finds, for each stopping time τ, an e-variable E_τ^{RIPR}

for $\mathbf{P}$. To define E_τ^{RIPR}, we first choose a $\bar{Q}$ via the plug-in or mixture method, exactly like we did above for UI. Then we consider the set $\mathbf{W}$ of all probability distributions on $\mathbf{P}$, and for each $W \in \mathbf{W}$, we denote by P_W the distribution obtained by mixing the distributions in $\mathbf{P}$ with W.

Extending results of Li (1999); Li and Barron (2000), Grünwald, De Heide and Koolen (2023, Theorem 1) show that, provided the infimum below is finite, for every $\tau \in \mathcal{T}$, there exists a unique measure P^τ for X^τ satisfying

$$D(\bar{Q}^\tau \| P^\tau) = \inf_{W \in \mathbf{W}} D(\bar{Q}^\tau \| P^\tau_W), \tag{15}$$

where $D(\cdot\|\cdot)$ is Kullback-Leibler divergence and $\bar{Q}^\tau$ (resp. P^τ_W) is $\bar{Q}^\tau$'s (resp. P^τ_W's) marginal for X^τ. Further, P^τ has the following nontrivial property: defining

$$E_\tau^{\text{RIPR}} := \bar{q}(X^\tau)/p^\tau(X^\tau),$$

where p^τ is the density of P^τ, and $\bar{q}$ is the density of $\bar{Q}$, E_τ^{RIPR} is an e-variable; it is even the GRO e-variable relative to τ, maximizing (12) over all e-variables that can be written as a measurable function of X^τ. P^τ is called the *reverse information projection* of $\bar{Q}$ onto $\mathbf{P}$. In some cases (e.g. Section 4.2, 4.1) it is easy to calculate, in others (Section 4.3) it is not. In general, it is a sub-probability measure, i.e. p^τ may integrate to less than one; but in all cases of practical interest we have encountered so far, P^τ is a probability distribution. In particular, because of the convexity of KL divergence and the set of mixtures of $\mathbf{P}$, the infimum is often achieved by some $W \in \mathbf{W}$ and then $P^\tau = P^\tau_W$. The sequence $(E_t^{\text{RIPR}})_{t\geq 1}$ is adapted to $\mathbf{F}$. Sometimes it is an e-process; sometimes not. More precisely:

1. For some $\Pi, \mathbf{P}, \mathbf{Q}$ (e.g. in Section 4.1), $(E_t^{\text{RIPR}})_{t\geq 1}$ is an e-process relative to an appropriately chosen $\bar{Q}$. It then dominates UI when using the same mixture over $\mathbf{Q}$: $E_t^{\text{UI}} \leq E_t^{\text{RIPR}}$, since they have the same numerator, but UI maximizes denominator likelihood.
2. For some $\Pi, \mathbf{P}, \mathbf{Q}$, $(E_t^{\text{RIPR}})_{t\geq 1}$ is an e-process when for $\bar{Q}$ we take any fixed $Q \in \mathbf{Q}$, but not when we take plug-in or mixture $\bar{Q}$ that 'learns'. This happens in Section 4.2. To handle composite $\mathbf{Q}$ we must then combine RIPr with the method of mixtures used as in Section 3.3.3.
3. For some $\Pi, \mathbf{P}, \mathbf{Q}$, $(E_t^{\text{RIPR}})_{t\geq 1}$ does not define an e-process. This happens in the example of Section 4.3. In this case, E_τ^{RIPR} can still be used to represent evidence in a study that stops at τ and can be multiplied with other e-variables in a meta-analysis setting (Section 6.1). But if we want to engage freely in optional stopping, we must use other methods, such as UI or the "sequential" version of RIPr illustrated in Section 4.3.

When $\mathbf{Q}$ is composite but we lack prior knowledge to justify a mixing distribution, one could use the GRAPA method from Section 3.2.5 that employs a smoothed empirical distribution. Alternatively, we may be able to obtain REGROW e-variables using RIPr, provided that we reformulate REGROW (13) as

$$\inf_{Q \in \mathbf{Q}} \mathbb{E}_Q \left(\log E_\tau - \log E_\tau^{\text{RIPR}(Q)} \right), \tag{16}$$

where $E_\tau^{\text{RIPR}(Q)}$ is the RIPr of Q onto $\mathbf{P}$. We hope to find a single e-variable E that approximately maximizes (16) simultaneously for every τ. In principle this task is well-defined even if the RIPr e-variables do not define an e-process; but it is much simplified if they are, for then we can apply the method of mixtures again to find an E_τ that approximately maximizes (16).

3.3.3 Mixing E-Processes. In all nonparametric, and some parametric cases, a natural way to proceed is to first construct a parameterized collection of e-processes $\{E^\lambda : \lambda \in \Lambda\}$. We then need to come up with a final e-process to use in practice. For this, we can use the method of mixtures again, but now by putting a distribution R on the space Λ and creating the new e-process E^R with, for each τ, $E_\tau^R := \int E_\tau^\lambda dR(\lambda)$, thus applying the method of mixtures directly to e-processes rather than to the alternative hypothesis $\mathbf{Q}$, as we did above when introducing UI and RIPr. In general these approaches are different, as will be illustrated in the parametric examples below: Section 4.1 mixes over $\mathbf{Q}$, Sections 4.2 and 4.3 (and all of Section 5) mix over a collection of e-processes.

4. PARAMETRIC EXAMPLES

Examples of test martingales and e-processes for simple nulls abound in the Bayesian literature, since every Bayes factor for a simple null also defines a test martingale. Further, as pointed out by Darling and Robbins (1968), if X_i are iid from P, and P has a finite MGF, meaning $\Phi(\lambda) := \mathbb{E}_P[\exp(\lambda X_i)] < \infty$, then $\exp(\lambda \sum_{i\leq t} X_i)\Phi(\lambda)^{-t}$ forms a test martingale for P. Thus, we emphasize examples with composite nulls. The examples of Section 4.1–4.3 are all implemented in the R package `safestats` on CRAN (Turner et al., 2022).

4.1 t-test, Regression, General Group Invariance

TODO PETER Consider the following version of the t-test: according to the null, the X_t are iid $\sim N(\delta_0\sigma, \sigma)$ for some given effect size δ_0; according to the alternative, they are iid $N(\delta_1\sigma, \sigma)$ for effect size δ_1. Under both null and alternative, the nuisance parameter σ is unknown, making the hypotheses composite. We coarsen the original process to $V_1, V_2, \ldots$, where $V_i := X_i/|X_1|$; of course, $|V_1| = 1$. Under the null, $(V_t)_t$ has the distribution P_{δ_0} that does not depend on the variance; similarly, it has a distribution P_{δ_1} that is the same under all distributions in the alternative. So by considering $(V_t)_t$ instead of $(X_t)_t$ we reduce the problem to a simple-vs-simple test as in Section 3.1, and the likelihood ratio $E_t := p_{\delta_1}(V^t)/p_{\delta_0}(V^t)$ is a test martingale for the null relative to a coarsened filtration. Essentially the same likelihood ratio was proposed by Rushton (1950) for classical sequential testing. Cox (1952) noted (using different terminology) that it can be rewritten as a Bayes factor applied to the original data under the improper right-Haar prior, $w(\sigma) = 1/\sigma$, i.e.

$$E_t = \frac{\int_{\sigma>0} p_{\delta_1\sigma,\sigma}(X^t)w(\sigma)d\sigma}{\int_{\sigma>0} p_{\delta_0\sigma,\sigma}(X^t)w(\sigma)d\sigma}, \tag{17}$$

where $p_{\mu,\sigma}$ denotes the density of a $N(\mu,\sigma)$ distribution. Lai (1976) also noted the equality (17) and first proposed to use E_t in an anytime-valid context, and even inverted the test to yield a closed-form confidence sequence for a Gaussian mean with unknown variance.

More recently Grünwald, De Heide and Koolen (2023) showed that E_t is the REGROW e-variable (16) for this problem (with θ set to σ and $E_\tau^{\text{RIPR}(\sigma)}$ equal to the GRO e-variable for testing simple alternative $\{P_{\delta_1\sigma,\sigma}\}$ vs. composite null $\{P_{\delta_0\sigma,\sigma} : \sigma > 0\}$, and τ arbitrary), implying that the proposed e-process is in fact optimal in a strong sense. They also demonstrate related growth optimality properties for the case that the defining constraint in either H_0 or H_1 or both is replaced by an inequality, i.e. H_0 expresses $\delta \leq \delta_0$ and/or H_1 expresses $\delta \geq \delta_1$, and for the case that, under H_0, $\delta = 0$ while under H_1, δ is equipped with a prior distribution. In the latter case, if a heavy-tailed prior is used, the marginal of (17) with respect to this prior coincides with the *Bayesian t-test* (Jeffreys, 1961; Rouder et al., 2009) which is thereby seen to have an e-process interpretation and to provide Type-I error safety. Thus, in this composite setting (and in contrast to the 2 × 2 and logrank examples below), test martingales overlap with a specific popular type of Bayes factors.

Pérez-Ortiz et al. (2022) extend these insights to the general setting of H_0 and H_1 that share nuisance parameters expressing a group invariance. For all such problems, one can determine a *sequence of maximal invariants*—a special process adapted to $\mathbf{F}$ that has the same distribution P_{δ_1} under all distributions in the alternative and the same distribution P_{δ_0} under all distributions in the null. In the t-test case, this is the process $(V^t)_t$ given above. Then the likelihood ratio of the maximal invariants always defines an e-process; and—the main contribution of Pérez-Ortiz et al. (2022)—under some conditions on the group, the most important of which is *amenability*, this likelihood ratio is, for each fixed n, the REGROW e-variable relative to H_0 and H_1. One can show that all required conditions, including amenability, hold for scale invariance such as in the t-test above, but also location and rotation invariance. By considering the affine group, one can handle linear regression problems with Gaussian errors and with the null having one of the parameters (the 'control') set to 0.

4.2 Parametric 2-Sample Tests; 2 × 2 Tables

Consider data streams $Y_{1,a}, Y_{2,a}, \ldots$ and $Y_{1,b}, Y_{2,b}, \ldots$, all $Y_{i,a}$ and $Y_{i,b}$ taking values in the same space $\mathcal{Y}$. We fix a statistical model $\{P^\circ_\theta : \theta \in \Theta\}$ of distributions on $\mathcal{Y}$. Our set Π is then given as $\Pi = \{P_{\theta_a,\theta_b} : (\theta_a, \theta_b) \in \Theta^2\}$, i.e., all distributions on Ω for which $Y_{i,a}$ are iid $\sim P^\circ_{\theta_a}$ and $Y_{i,b}$ are iid $\sim P^\circ_{\theta_b}$ for some $(\theta_a, \theta_b) \in \Theta^2$. In the simplest setting, data is further constrained to come in 'blocks' $X_1, X_2, \ldots$, each block X_j containing $n_a \geq 1$ elements of stream a and $n_b \geq 1$ elements of stream b, and we want to test the null hypothesis that $\theta_a = \theta_b$, i.e. $\mathbf{P} = \{P_{\theta,\theta} : \theta \in \Theta\}$. We first consider the case of testing this composite null vs. a simple alternative $\mathbf{Q} = \{P_{\theta_a,\theta_b}\}$. Turner, Ly and Grünwald (2021) design a 1-round e-variable $S_j^{\theta_a,\theta_b}$ for a single block of data $X_j = (Y^{(j)}_{1,a}, \ldots, Y^{(j)}_{n_a,a}, Y^{(j)}_{1,b}, \ldots, Y^{(j)}_{n_b,b})$ with n_a outcomes in group a and n_b outcomes in group b, defined as

$$S_j^{\theta_a,\theta_b} = \frac{p_{\theta_a}(Y^{(j)}_{1,a}, \ldots, Y^{(j)}_{n_a,a}) \cdot p_{\theta_b}(Y^{(j)}_{1,b}, \ldots, Y^{(j)}_{n_b,b})}{\prod_{g\in\{a,b\}, i=1\ldots n_g} \left(\frac{n_a}{n_a+n_b} p_{\theta_a}(Y^{(j)}_{i,g}) + \frac{n_b}{n_a+n_b} p_{\theta_b}(Y^{(j)}_{i,g})\right)}.$$

If one observes iid data blocks $X_1, X_2, \ldots, X_t$ (each X_j represents a full block), one can calculate $S_j^{\theta_a,\theta_b}$ for each block $j = 1, \ldots, t$ and set $E_t^{\theta_a,\theta_b} := \prod_{j=1}^t S_j^{\theta_a,\theta_b}$. Then E^{θ_a,θ_b} is a test martingale. One can accommodate a composite alternative $\mathbf{Q} = \{P_{\theta_a,\theta_b} : (\theta_a, \theta_b) \in \Lambda\}, \Lambda \subset \Theta^2$, by learning (θ_a, θ_b) using the method of mixtures as in Section 3.3.3 with λ set to (θ_a, θ_b). One can extend this picture in variouis ways, e.g. letting group sizes vary over time and using preceding data to determine the next n_a and n_b.

Here we concentrate on the special case that $\mathcal{Y} = \{0, 1\}$, in which the e-processes embody a sequential version of the 2 × 2-table contingency test, and for each τ, $E_\tau^{\theta_a,\theta_b}$ turns out to be equal to $E_\tau^{\text{RIPR}(P_{\theta_a,\theta_b})}$ as obtained by (15), and is thus GRO relative to simple alternative $\mathbf{Q} = \{P_{\theta_a,\theta_b}\}$. One can now determine a prior R as in Section 3.3.3 that approximately optimizes the REGROW criterion (16), providing strong optimality guarantees for the constructed e-process. Turner, Ly and Grünwald (2021) determine the prior that optimizes (16) among all beta-priors, and find that it behaves excellently in practice. Turner and Grünwald (2023) then give the corresponding confidence sequence for various notions of effect size such as absolute difference $\phi(\theta_a, \theta_b) = \theta_b - \theta_a$, log-relative risk $\phi(\theta_a, \theta_b) = \log(\theta_b/\theta_a)$ and the log-odds ratio $\phi(\theta_a, \theta_b) = \log(1 - \theta_a)\theta_b/(\theta_a(1 - \theta_b))$.

4.3 Logrank Test and Cox Regression

Ter Schure et al. (2021) provide efficiently computable e-variables and test martingales for the logrank test, a work-horse of medical statistics. Here we describe the test martingale they derive for the more general setting of the Cox regression model with covariates, for which computationally efficient implementation remains a work in progress. One starts with m subjects, partitioned into a *treatment* and a *control* group; for example, one wants to test a COVID vaccine; at time 0 all m_1 subjects in the treatment group get the vaccine and all $m - m_0$ subjects in the control get a placebo. Each subject $j \in [m]$ (where $[m] := \{1, \dots, m\}$) is associated with a $(d+1)$-dimensional covariate vector $z_j = (z_{j,0}, z_{j,1}, \dots, z_{j,d})$ where the special covariate $z_{j,0} \in \{0,1\}$ denotes whether the subject is in the treatment ($z_{j,0} = 1$) or control ($z_{j,0} = 0$) group. Data $X_1, X_2, \dots$ come in sequentially, in the form of *events*. In our example, events $X_1 = j_1, X_2 = j_2$ would represent 'the first to get COVID was subject j_1; the second was j_2'. The *risk set* $\mathcal{R}_n$ at time n is the set of subjects that did not have an event yet; i.e. if $X_1 = j_1, \dots, X_n = j_n$, then $\mathcal{R}_{n+1} = [m] \setminus \{j_1, \dots, j_n\}$. Cox's celebrated model rests on the *proportional hazards* assumption. This assumption implies that, conditional on the first $n-1$ events, the probability of the n-th event happening to any subject j that is still in $\mathcal{R}_n$ is given by

$$(18) \qquad P_\beta(X_n = j \mid X^{n-1} = j^{n-1}) = \frac{\exp(\beta^T z_j)}{\sum_{j' \in \mathcal{R}_n} \exp(\beta^T_{j'} z_{j'})}$$

for some parameter vector $\beta = (\beta_0, \dots, \beta_d)$. We would like to test the null hypothesis that $\beta_0 = 0$ (no effect of treatment) against alternative $\beta_0 \leq -\delta$ or $\beta_0 \geq \delta$ for some $\delta > 0$ (in our example, we would take alternative $\beta_0 \leq -\delta$ indicating that vaccination reduces the chance of getting COVID). Proceeding as in the previous example, we start with a simple alternative, i.e. pretend that we know that *if* $\beta_0 \neq 0$, then $\beta = \beta_{(1)}$ for some fixed $\beta_{(1)} \in \mathbf{R}^{d+1}$. We can then determine, for each n, the *single-round* GRO e-variable $S_n^{\beta_{(1)}}$ for the sample-size-1 outcome X_n under alternative given by (18) with β set to $\beta_{(1)}$. The distribution in (18) being dependent on the past, the definition of this e-variable will also depend on the past: $S_n^{\beta_{(1)}} := p_{\beta_{(1)}}(X_n \mid X^{n-1}) / p_{W \leftsquigarrow \beta_{(1)}}(X_n \mid X^{n-1})$ where $W \leftsquigarrow \beta_{(1)}$ is a distribution on the null model parameter space $\{\beta \in \mathbb{R}^{d+1} : \beta_0 = 0\}$ such that, conditionally on each x^{n-1}, $P_{W \leftsquigarrow \beta_{(1)}}(X_n \mid X^{n-1} = x^{n-1})$ is the RIPr of $P_{\beta_{(1)}}(X_n \mid X^{n-1} = x^{n-1})$ as given by (15). Since the sample space (i.e. the risk set) on which (18) is defined is finite, we can guarantee that the RIPr is a finite mixture, i.e. $W \leftsquigarrow \beta_{(1)}$ is a distribution with finite support on $\{\beta \in \mathbb{R}^{d+1} : \beta_0 = 0\}$. Thus, at least in principle, S_n^{β} can be found by numerical optimization methods. This gives us a process $S^{\beta_{(1)}}$ of single-round past- conditional e-variables as in (11), i.e. each $S_n^{\beta_{(1)}}$ is a function of X^n, and, under every distribution in the null, $\mathbb{E}[S_n^{\beta_{(1)}} \mid X^{n-1}] \leq 1$, so that their running product is a test martingale, with good growth behaviour under the simple alternative indexed by $\beta_{(1)}$. We can apply the method of mixtures as in Section 3.3.3, with λ in the role of $\beta_{(1)}$, putting a prior distribution R on $\beta_{(1)}$, to go from a 'simple' to composite alternative. We note that Cox's model is highly nonparametric—it assumes continuous time. By appropriately coarsening the data Cox arrived at the 'partial' likelihood (18) which allows us to treat the problem as if it were parametric. Thus when we write 'simple' above we really mean a large set of continuous time processes that all satisfy (18) for a single parameter vector β.

Ter Schure et al. (2021) have not yet found any method to calculate or approximate the prior $W \leftsquigarrow \beta_{(1)}$ (and hence the required e-variables $S_n^{(\beta_1)}$) for this problem if $d > 0$. Yet the situation simplifies dramatically if $d = 0$, i.e. the only covariate is the binary treatment/control dichotomy. For that situation, under distribution P_β with $\beta = \beta_0 \in \mathbb{R}$, $P_\beta(X_n \mid X^{n-1} = j^{n-1}, z_{X_n} = g)$ is uniform for both groups $g \in \{0,1\}$: all remaining subjects in group g have the same probability of an event, which only depends on g. The probability of X_n being in group 1 (the next event is in the treatment group) given X^{n-1} becomes $n_1 \exp(\beta) / (n_1 \exp(\beta) + n_0)$, where n_g is the number of subjects in the risk set that are in group g. under the null ($\beta = 0$) this is simply a Bernoulli determined by the relative group size. The resulting test can then be interpreted as a sequential version of the classical logrank test (a similar test martingale was independently, and without reference to survival analysis, proposed by Lindon and Malek (2020)). Ter Schure et al. (2021) give simulations comparing it to various standard group sequential/α-spending approaches (Section 8.1.2).

5. NONPARAMETRIC EXAMPLES

We begin with case studies that illustrate how one might build (A) test martingales for composite nonparametric $\mathbf{P}$ despite there being no common reference measure, (B) test supermartingales for $\mathbf{P}$ when no test martingales exist, (C) e-processes for $\mathbf{P}$ when no test supermartingales exist, and (D) confidence sequences for functionals using submartingales in reversed time instead of test martingales.

5.1 Estimating Sub-Gaussian Means (Case B)

A distribution P for real-valued $X_1, X_2, \dots$ is *sub-Gaussian* with parameter $\sigma > 0$ if

$$\forall \lambda, i : \mathbb{E}_P[\exp(\lambda(X_i - \mu_i)) \mid \mathbf{F}_{i-1}] \leq \exp(\lambda^2 \sigma^2 / 2),$$

where $\mu_i := \mathbb{E}_P[X_i \mid \mathbf{F}_{i-1}]$. Fixing σ, let $\mathbf{G}^\mu$ be the set of sub-Gaussian distributions with parameter σ and $\mu_i = \mu$ for all i. Set $\mathbf{P} := \{\mathbf{G}^\mu\}_{\mu \in \mathbb{R}}$. This is a nonparametric generalization of Gaussianity; the conditional mean is constant and the conditional moment generating function is no larger than that of a Gaussian with variance σ. Darling and Robbins (1968) effectively constructed CSs for μ. They first observed that for any $\lambda \in \mathbb{R}$,

$$(19) \qquad M_t^\mu(\lambda) := \exp\left(\lambda \sum_{i \leq t} (X_i - \mu) - \tfrac{\lambda^2}{2} \sigma^2 t\right)$$

is a test supermartingale for $\mathbf{G}^{\mu}$. Setting $Y_t := \sum_{i\le t} X_i$, and choosing a centered Gaussian with variance ρ^2 as a mixing distribution F, they define

$$M_t^{\mu} := \int M_t^{\mu}(\lambda)dF(\lambda) = \frac{\exp(\frac{\rho^2(Y_t - t\mu)^2}{2(t\sigma^2\rho^2+1)})}{\sqrt{t\rho^2\sigma^2+1}}.$$

Since the supermartingale property is closed under averaging, M_t^{μ} is also a test supermartingale for $\mathbf{G}^{\mu}$. It grows exponentially fast under any $P \in \mathbf{P}^{\theta}$ for any $\theta \neq \mu$, and it grows faster for θ far from μ; thus, the evidence automatically adapts to the difficulty of the testing problem. This type of adaptivity is a commonly observed benefit of the method of mixtures, if appropriately employed.

Using the inversion (10), we find that

$$\text{(20)} \qquad \frac{Y_t}{t} \pm \sigma \sqrt{\frac{(t\rho^2+1)}{t^2\rho^2}\log((t\rho^2+1)/\alpha^2)}$$

is a CS for μ. If μ_i differs for each i, an identical argument shows that (20) is a CS for the running mean $\sum_{i=1}^{t} \mu_i/t$. The general scaling of $\sigma t^{-1/2}\sqrt{\log t + \log \alpha^{-1}}$ is expected when using mixture distributions that are continuous around the origin (Howard et al., 2021, Prop. 2). The $\sqrt{\log t}$ can be changed to $\sqrt{\log\log t}$ at the expense of other constants using mixture distributions that are unbounded at the origin; see Howard et al. (2021, Eq. (1)).

Waudby-Smith and Ramdas (2023) observed that

$$\exp\left(\sum_{i\le t}\left(\lambda_i(X_i-\mu) - \tfrac{\lambda_i^2}{2}\sigma^2\right)\right)$$

is also a test supermartingale for $\mathbf{G}^{\mu}$, whenever λ_i is predictable. They invert plug-in test supermartingales of this form to yield efficient CSs.

5.2 Heavy-Tailed Mean Estimation (Case B)

For a given $\sigma > 0$, let $\mathbf{V}^{\mu}$ be the set of distributions on $\mathbb{R}^{\infty}$ that yield observations with conditional mean μ and conditional variance bounded above by σ^2. The set $\mathbf{V}^{\mu}$ contains $\mathbf{G}^{\mu}$ and is much larger. A distribution in $\mathbf{V}^{\mu}$ may be very heavy tailed, with no more than the first two moments. Inspired by the seminal nonsequential work of Catoni (2012), one can prove that

$$L_t^{\mu} := \exp\left(\sum_{i\le t}\varphi(\lambda(X_i-\mu)) - \tfrac{\lambda^2}{2}\sigma^2 t\right)$$

is a test supermartingale for $\mathbf{V}^{\mu}$, where $\varphi(x)$ equals $\log(1 + x + x^2/2)$ if $x \ge 0$ and $-\log(1 - x + x^2/2)$ if $x < 0$. Wang and Ramdas (2023) use the plug-in and inversion techniques discussed in the previous cases to derive a CS for μ. Somewhat surprisingly, these CSs for μ (assuming $P \in \mathbf{V}^{\mu}$) appear, visually, almost identical to the sub-Gaussian CSs one gets when assuming $P \in \mathbf{G}^{\mu}$. In other words, for mean estimation, the sub-Gaussian assumption — which is very common in machine learning (in the multi-armed bandit literature, for example) — can be relaxed with almost no practical consequence. Wang and Ramdas (2023) also derive simple extensions for the case when the p-th moment is finite for $p < 2$.

There are other known test supermartingales for this setting, for example by Dubins and Savage (1965) and Delyon (2009, Proposition 12) but Wang and Ramdas (2023) find these to be less powerful.

5.3 Variance-Adaptive Estimation of Bounded Means (Case A)

In the previous two examples, the sub-Gaussian parameter or the variance bound σ must be provided (or an upper bound must be guessed) in advance by the statistician, since it is provably impossible to learn σ from the data itself. Unfortunately, neither CS adapts to the unknown variance by yielding a tighter CS for lower variance data (and it is likely impossible to design ones that can); if $\mathbb{E}[(X_i-\mu)^2|\mathbf{F}_{i-1}] \ll \sigma^2$, the CS will still depend on the provided conservative value σ. These impossibilities are due to the unbounded nature of $\mathbf{G}$ and $\mathbf{V}$, allowing some distributions to have very small mass far away from the origin. For the subclass of bounded random variables, however, variance-adaptive mean estimation is feasible.

Let $\mathbf{B}^{\mu}$ denote the set of distributions P on $[0,1]^{\infty}$ such that $\mathbb{E}_P[X_i|\mathbf{F}_{i-1}] = \mu$. Howard et al. (2021) prove that for any $\lambda \in [-1,1]$, and any predictable $\hat{\mu}_i \in \mathbf{F}_{i-1}$,

$$\text{(21)} \quad N_t^{\mu}(\lambda) := \exp\left(\lambda\sum_{i\le t}(X_i-\mu) - \psi(\lambda)\sum_{i\le t}(X_i-\hat{\mu}_i)^2\right),$$

where $\psi(\lambda) := -\log(1-\lambda) - \lambda$, is a test supermartingale for $\mathbf{B}^{\mu}$. Because ψ is the logarithm of the moment generating function (MGF) of a centered unit-rate exponential distribution, we call N^{μ} a subexponential supermartingale. As $\lambda \to 0$, $\psi(\lambda)$ behaves like the Gaussian log-MGF $\lambda^2/2$, but unlike the sub-Gaussian supermartingale in (19), we can employ a fully empirical variance term in (21). This generalizes a result of Fan, Grama and Liu (2015), who effectively proved the same claim with $\hat{\mu}_i := 0$ for all i. The extension to predictable $\hat{\mu}_i$, which requires some additional tricky algebra, is very useful in lowering the empirical variance.

As before, we need to mix over the tuning parameter λ. The gamma family provides suitable mixing distributions, because they are conjugate to the exponential, leading to a closed form mixture supermartingale. The resulting CS was developed in Howard et al. (2021).

Waudby-Smith and Ramdas (2023) note that for predictable $\lambda_i \in \mathbf{F}_{i-1}$,

$$N_t^{\mu} := \exp\left(\sum_{i\le t}\lambda_i(X_i-\mu) - \sum_{i\le t}\psi(\lambda_i)(X_i-\hat{\mu}_i)^2\right)$$

is also a subexponential "plug-in" test supermartingale that can be tuned to closely mimic the earlier mixture supermartingale. This means that

$$\frac{\sum_{i\le t}\lambda_i X_i}{\sum_{i\le t}\lambda_i} \pm \frac{\log(2/\alpha) + \sum_{i\le t}\psi(\lambda_i)(X_i - \hat{\mu}_i)^2}{\sum_{i\le t}\lambda_i}$$

is a $(1-\alpha)$-CS for μ.

The preceding techniques are interesting because they can be used even when the observations are not bounded. But for the bounded model $\cup_{\mu\in\mathbb{R}}\mathbf{B}^\mu$, the most statistically powerful way to derive a CS for μ is to use plug-in test martingales for $\mathbf{B}^\mu$ of the form

$$K_t^\mu := \prod_{i=1}^{t}(1 + \lambda_i^\mu(X_i - \mu)), \tag{22}$$

where λ_i^μ is a predictable process indexed by μ. As before, $C_t := \{\mu : K_t^\mu < 1/\alpha\}$ is a $(1-\alpha)$-CS for μ. The λ_i^μ are naturally interpreted as bets on the X_i; they must be predictable because a bet on X_i must be made before seeing X_i. This idea was suggested by Hendriks (2018), and was independently proposed and studied in more depth by Waudby-Smith and Ramdas (2023), who derive betting strategies that are adaptive to the underlying distribution P, in particular to its mean and variance, establishing connections to the Chernoff method, empirical and dual likelihood, and other parts of the literature. Followup work by Orabona and Jun (2021) derives other betting strategies via connections to Thomas Cover's universal portfolios. None of these strategies uniformly dominates any other, and the resulting CSs are in general not comparable (some may be tighter earlier but looser later, etc.).

5.4 Testing Symmetry (Case A)

Let $\mathbf{P}$ be the set of distributions on $\mathbb{R}^\infty$ such that X_t and $-X_t$ have the same distribution given $\mathbf{F}_{t-1}$, for every $t \ge 1$. Extending an older result by Efron (1969), de la Peña (1999, Lemma 6.1) establishes that for any $\lambda \in \mathbb{R}$,

$$R_t(\lambda) := \exp\left(\lambda \sum_{i\le t} X_i - \frac{\lambda^2}{2}\sum_{i\le t} X_i^2\right) \tag{23}$$

is a test supermartingale for $\mathbf{P}$. Notice again the fully empirical variance term, as in (21); these are also called self-normalized processes. As before, one can mix over λ or use the plug-in technique described earlier.

Recently, Ramdas et al. (2020) proved that R_t is inadmissible for testing symmetry by constructing a test *martingale* R_t^o for $\mathbf{P}$ that is always at least as large as R_t, and typically larger. In fact, M is a test martingale for $\mathbf{P}$ (and is admissible) if and only if the unit bets B_t at time t in (4) are nonnegative and predictable, and $B_t - 1$ is an odd function of X_t. Note that the unit bet underlying (23) takes the form $\exp(\lambda x - \lambda^2 x^2/2)$, which does not yield an odd function on subtracting one; rectifying this yields the improved and admissible test martingale R_t^o.

5.5 Testing Exchangeability and Log-Concavity (Case C)

In the previous example, $\mathbf{P}$ is a very rich, nonparametric class of distributions (discrete and continuous, light and heavy tailed, etc.) with no common dominating measure. Being able to find a single (nonconstant) process that is simultaneously a test martingale for every P in $\mathbf{P}$ is quite atypical, and it can only occur for very structured problems. (The same atypical situation also occurred with $\mathbf{B}^\mu$ in the bounded case.) For example, there is no nontrivial test martingale for $\mathbf{G}^\mu$, the sub-Gaussian class discussed earlier; nevertheless, we did exhibit a test supermartingale. It turns out that even this is atypical: a rather special structure is required for a (nonmonotonic) test supermartingale to exist.

Ramdas et al. (2022) study the seemingly simple problem of testing if a binary sequence is exchangeable, and find that *no nontrivial test supermartingale exists* (in the original filtration), but they exhibit a nontrivial and powerful e-process based on universal inference.

Remarkably, Vovk (2021) demonstrates that by *shrinking the filtration* to include only conformal p-values, it is once again possible to design nontrivial test martingales, even though none exist in the richer data filtration. Vovk's method works for general observation spaces, but in the binary case, experiments by Vovk, Nouretdinov and Gammerman (2021) demonstrate that it is not as powerful as the aforementioned e-process.

Another relevant example is that of testing log-concavity. Let $\mathbf{L}_d$ denote the set of distributions P on $\mathbb{R}^d$ with Lebesgue densities p such that $\log p(x)$ is concave in x. $\mathbf{L}_d$ is a nonparametric class that contains all Gaussian, logistic, exponential and Laplace distributions, as well as uniform distributions on any convex set. It is possible to prove with some effort that there is no test supermartingale for $\mathbf{L}_d$. However, the universal inference approach yields a powerful e-process for $\mathbf{L}_d$; see Dunn et al. (2021).

Of course, there are problems for which even no nontrivial e-process exists and testing those nulls is futile; see Ruf et al. (2022) for examples.

5.6 Estimating Convex Functionals and Divergences by Reversing Time (Case D)

Consider a set of probability distributions Π that is closed under convex combinations. A functional $\phi : \Pi \mapsto \mathbb{R}_{\ge 0}$ is called convex if $\phi(aP + (1-a)Q) \le a\phi(P) + (1-a)\phi(Q)$ for any $P, Q \in \Pi$ and $a \in [0,1]$. Classic examples are the entropy and the mean. Similarly, a divergence $D : \Pi\times\Pi \mapsto \mathbb{R}_{\ge 0}$ is called convex if $D(aP+(1-a)P', aQ+(1-a)Q') \le aD(P,Q) + (1-a)D(P',Q')$. Examples include the total variation distance, Kullback-Leibler divergence, kernel maximum mean discrepancy, Kolmogorov-Smirnov distance, Wasserstein distance or any integral probability metric or f-divergence.

Suppose $X_1, X_2, \dots, X_t, \dots \sim P$ and let P_t denote the empirical distribution of X^t. The exchangeable filtration $\mathcal{E}_t$ is the *decreasing* filtration given by $\mathcal{E}_t = \sigma(P_t, X_{t+1}, X_{t+2} \dots)$; in words: $X_{t+1}, X_{t+2} \dots$ are known perfectly, but the order of $X_1, X_2, \dots, X_t$ is forgotten. Manole and Ramdas (2023) derive a curious property: for any convex functional ϕ, the process $(\phi(P_t))_{t\geq 1}$ is a *reverse* submartingale with respect to the exchangeable filtration. (An analogous statement also applies to divergences.)

Recall that a reverse submartingale is a submartingale when time is reversed and the process is viewed from time ∞ to zero. Reverse submartingales behave somewhat like forward supermartingales: their expectations are decreasing as time increases. Nonnegative reverse submartingales behave like test supermartingales in that there exists a *reverse* Ville's inequality, with an identical statement to the forward Ville's inequality. Manole and Ramdas (2023) use this to derive confidence sequences for (say) the entropy of a distribution, as well as for divergences between pairs of distributions in quite some generality. The same technique also allows the authors to derive the first tight CSs (in their dependence on sample size, dimension, etc.) for suprema of Gaussian processes, Rademacher complexities, U-statistics, quantile functions, and several other interesting objects.

A game-theoretic interpretation of nonnegative reverse submartingales remains unknown, as does a game-theoretic derivation of the above CSs.

5.7 Sequential Change Detection

On observing a stream of data, the problem of sequential change detection can be seen as an extension of sequential testing: either all the data is from some $P \in \mathbf{P}$, or at some time ν, it switches from P to some $Q \in \mathbf{Q}$. If there is indeed a change, we would like to stop as quickly as possible and proclaim a change (call this time τ^*). Measures of the performance of a change detection procedure include average run length (ARL, also called frequency of false alarms) and average detection delay. These are respectively defined as $\inf_{P\in\mathbf{P}} \mathbb{E}_P[\tau^*]$, and $\sup_{P\in\mathbf{P},\nu>0,Q\in\mathbf{Q}} \mathbb{E}_{P,\nu,Q}[\tau^* - \nu \mid \tau^* > \nu]$, where the subscript P, ν, Q means that the data come from P up through time ν and from Q after ν. We would like the former to be as large as possible with the latter being as small as possible.

Extending Volkhonskiy et al. (2017), who use conformal test martingales to detect deviations from exchangeability of the X_i's, Shin, Ramdas and Rinaldo (2022) describe a general nonparametric game-theoretic framework for change detection. They define an *e-detector* for $\mathbf{P}$ to be a nonnegative process M such that

$$\mathbb{E}_P[M_\tau] \leq \mathbb{E}_P[\tau] \text{ for every } \tau \in \mathcal{T} \text{ and } P \in \mathbf{P}.$$

(Like an e-process, the definition only depends on $\mathbf{P}$ but we measure its quality relative to a post-change class $\mathbf{Q}$). The authors show that if one can construct an e-process for $\mathbf{P}$, then one can define an e-detector for $\mathbf{P}$ by summing e-processes started at consecutive times. Formally, $M_t := \sum_{i\leq t} A_i$ is an e-detector for $\mathbf{P}$, where A_i is an e-process for $\mathbf{P}$ that depends only on $X_i, X_{i+1}, \dots$. Game-theoretically, M_t is the wealth of a gambler who injects an extra dollar into the game at each time, and uses it to bet against $\mathbf{P}$.

The above definition and construction may appear mysterious, but yields methods with nontrivial properties. First, defining $\tau^* := \inf\{t \geq 1 : M_\tau \geq 1/\alpha\}$, one can prove that the ARL is at least $1/\alpha$. Second, in certain parametric settings, if there is indeed a changepoint, then one can design e-detectors such that the detection delay is near-optimal in a particular sense: even if P, Q were known in advance, the best possible detection delay for any method with ARL at least $1/\alpha$ scales like $\log(1/\alpha)/D(P\|Q)$, where (as before) D is the Kullback-Leibler divergence. An e-detector based on likelihood ratios recovers the famous Shiryaev-Roberts statistic and can adaptively achieve this optimal scaling (up to lower order terms) without knowing P, Q, by employing new mixture and plug-in approaches. Third, e-detectors can be built for many nonparametric $\mathbf{P}$ using (for example) the e-processes constructed earlier in this section. For many such nonparametric problems, e-detectors provide, as far as we know, the first changepoint procedures with provable ARL control.

5.8 *Asymptotic* Confidence Sequences and Sequential Causal Inference

The average treatment effect (ATE) is arguably the most popular estimand in causal inference, and one may ask if it is possible to estimate it sequentially in both randomized and observational settings. For brevity, we focus on the observational setting, where finite-sample inference is not possible due to unknown biases caused by confounding, but under suitable assumptions it is possible to design "doubly-robust" estimators for the ATE that have (nonsequential) asymptotic coverage guarantees. A suitable generalization of the concept of confidence sequences is required, because (by definition) CSs have finite-sample validity guarantees that we do not know how to achieve even at fixed sample sizes.

With such goals in mind, Waudby-Smith et al. (2021) define "asymptotic confidence sequences", which may sound paradoxical at first. They mirror an analogous definition of asymptotic CIs. Informally, a sequence of (measurable) sets $(C_t)_{t\geq 1}$ is called an asymptotic CS if there exists some unknown nonasymptotic CS $(D_t)_{t\geq 1}$, such that the measure of the symmetric difference between C_t and D_t almost surely vanishes faster than a $\sqrt{\log\log t/t}$ rate. Waudby-Smith et al. (2021) then derive a universality result: informally, as long as the data have more than two moments (an almost necessary condition for inference), a universal asymptotic CS is given by (20) but with σ replaced by an empirical variance $\hat{\sigma}_t$. This yields a time-uniform analog of the central limit theorem (CLT), and is established using certain strong approximation theorems for Brownian motion. The authors then construct doubly-robust asymptotic CSs for the ATE, yielding anytime versions of the corresponding CIs.

Beyond causal inference, asymptotic CSs can be used in a variety of other settings where CLT-based CIs are the norm in the offline setting. These include M-estimation and other semiparametric and nonparametric functional estimation problems; see also Pace and Salvan (2020) and Johari et al. (2022).

In complementary work, Duan, Ramdas and Wasserman (2022) develop test martingale versions of rank based tests like the Wilcoxon, Kruskal-Wallis, and Friedman tests. Batch versions of these tests are commonly used for testing the strong global null (of no treatment effect) in a randomized experiment with covariates.

5.9 Other Nonparametric Problems

5.9.1 A Compendium of Exponential Supermartingales. We have encountered several nonparametric test supermartingales of the form

$$\exp(\lambda \cdot (\text{sum}_t) - \psi(\lambda) \cdot (\text{variance}_t)).$$

Likelihood ratios for point nulls in exponential families also take this form. So in a very concrete sense, we have been generalizing the likelihood ratio to composite and nonparametric problems, and even to cases where there is no dominating measure to define likelihood ratios (most previous subsections). Just as likelihood ratios are fundamental objects for parametric inference, test (super)martingales and e-processes are fundamental objects for nonparametric inference. Howard et al. (2020) summarize a large literature on test supermartingales and e-processes of the above exponential form, in discrete and continuous time, for scalar-, vector- and matrix-valued observations, and under a variety of nonparametric conditions. We have mentioned only some examples.

5.9.2 Estimating Quantiles. Howard and Ramdas (2022) derive confidence sequences based on iid data for any prespecified quantile of an unknown probability distribution, improving on those derived by Darling and Robbins (1967). They also derive CSs for the entire cumulative distribution function (or quantile function), providing a time-uniform extension of the famous Dvoretzky-Kiefer-Wolfowitz inequality.

5.9.3 Two-Sample and Independence Testing. Here, we observe two samples and want to know if they have the same distribution, making no further assumptions. This is one of the best studied problems in statistics. Common methods in the offline setting include the univariate Kolmogorov-Smirnov test and the multivariate kernel maximum mean discrepancy, amongst many others. However, the literature on sequential nonparametric two-sample testing appears sparse: Balsubramani and Ramdas (2016) and Lhéritier and Cazals (2018) do use test (super)martingales, but the methods they proposed are not that powerful in practice (although Pandeva et al. (2022) report excellent results with an extension of the ideas in (Lhéritier and Cazals, 2018)). Shekhar and Ramdas (2021) describe a relatively general game-theoretic framework that provides the first sequential analog of large classes of offline nonparametric tests, and these perform rather well in practice. The evidence grows slowly for hard problems (when the two distributions are different but very similar) and quickly for easy ones (when the two distributions are very different), and it can be monitored and stopped adaptively. This is a major advantage over offline tests when the problem difficulty is not known in advance. Recently, the first sequential nonparametric independence testing framework was developed in Podkopaev et al. (2023), which allows random variables to lie in general spaces and handles non-i.i.d. settings.

5.9.4 Sampling Without Replacement (WoR). Another classical problem is that of estimating a mean when sampling WoR. Here we have a bag of N numbers $\{x_1, \dots, x_N\}$, say all in the range $[0, 1]$, and we wish to estimate their average $\mu := \sum_{i \le N} x_i/N$, or (say) to test if it is at most a half. The randomness arises from the WoR sampling process. Waudby-Smith and Ramdas (2020) construct powerful plug-in test supermartingales for testing such hypotheses (of the empirical Bernstein flavor in (21)), and invert them to construct CSs. Waudby-Smith and Ramdas (2023) designed more powerful test martingales of the form (22) and the resulting confidence sequences are state-of-the-art. These were then applied quite successfully towards election auditing by Waudby-Smith, Stark and Ramdas (2021) and more recently by Spertus and Stark (2022).

6. MULTIPLE HYPOTHESIS TESTING

6.1 Global Null Testing and Meta-Analysis

Based on test martingales, Ter Schure and Grünwald (2022) propose *ALL-IN* (*A*ny time *L*ive and *L*eading *IN*terim) meta-analysis. This meta-analysis can be updated *any time*, even after each new observation, while retaining type-I error guarantees. It is *live*: no need to specify in advance the times when you will look and reanalyze. And it can be the *leading* source of information for deciding whether individual studies should be initiated, stopped early, or expanded.

These authors illustrate the method for clinical trials involving time-to-event data, using a Gaussian approximation to Section 4.3's logrank test. Consider the case where each study tests the null hypothesis that some effect size δ (measuring, say, the efficacy of a medical treatment) is

0; extensions to CIs are possible via inversion. In the simplest case, the evidence for the i-th study is measured by a unit bet $S_{(i)}$; the null is always the same (a "global null"), but the alternative may change. For example, if the first study is based on the mixture method of Section 3.2, the mixing distribution for later studies might be updated using the outcomes of the studies so far or changed because the next study samples from a different population. The unit bets $S_{(1)}, S_{(2)}, \dots$ generated this way can be multiplied, so that the process E with $E_{(j)} := \prod_{i=1}^{j} S_{(i)}$ is a test martingale at the "meta-level", with individual outcomes replaced by entire studies. We can always keep initiating and adding new studies as we want at the time, deciding whether do so and choosing the unit bet for any new study in light of the outcomes of the previous studies.

In the terminology of Grünwald, De Heide and Koolen (2023), the method is safe under *optional continuation*. This is true when the study-level e-variables $S_{(j)}$ are produced by Section 3.3's RIPr, even when the RIPr does not give an e-process at the individual outcome level. The method is more flexible though when each study j is associated with an e-*process*. As Ter Schure and Grünwald (2022) show, it is then possible to interleave the studies—one may first observe some outcomes from study 1, then some from study 4, then some from study 1 again, etc., tracking the cumulative product of the e-variables resulting from each batch. Again, one can decide at any time to stop an individual study, initiate or change studies, or stop the meta-analysis all-together, while still retaining Type-I error guarantees throughout.

Without using the ALL-IN terminology, Duan et al. (2020) design several martingale methods to sequentially test a global null when each study ends with a p-value (instead of e-value) that is valid conditional on all past studies. These new methods can be seen as sequential analogs to several well known nonsequential p-value combination rules like Fisher's or Stouffer's. Alternatively, one could *calibrate* the p-values into e-values (calibrators are defined in Section 6.3) and multiply them as done above.

6.2 False Discovery Rate

The false discovery rate (FDR) is probably the most popular error metric in modern large-scale multiple testing. The BH procedure (Benjamini and Hochberg, 1995) is the standard procedure for controlling the FDR when working with p-values. Given a target FDR level α, it proclaims as discoveries the hypotheses corresponding to the k^* smallest p-values out of K, where

$$k^* := \max\{k \in \{1, \dots, K\} : p_{(k)} \le \alpha k/K\},$$

and $p_{(k)}$ represents the k-th smallest p-value. It is known to control the false discovery rate when the null p-values are independent of each other and of the non-nulls, as well as under a particular type of positive dependence known as PRDS (Benjamini and Yekutieli, 2001).

Wang and Ramdas (2022) define an analogous e-BH procedure, which rejects hypotheses corresponding to the k^* largest e-values, where

$$k^* := \max\{k \in \{1, \dots, K\} : e_{[k]} \ge K/(k\alpha)\},$$

and $e_{[k]}$ represents the k-th largest e-value. Surprisingly, this procedure controls the FDR at α under arbitrary dependence between all the e-values. In fact, the same fact is true if one picks any set of S e-values that are all larger than $K\alpha/S$, while an analogous result is known to not hold for p-values.

When both p-values and e-values are available for the same set of hypotheses (for example, from different datasets collected under different conditions), Ignatiadis, Wang and Ramdas (2022) define generalizations of the above procedures that use both sources of information. In particular, e-values can serve as unnormalized weights within standard FDR methods that use weighted p-values. The waiving of the need to normalize the weights (to sum to one) gives the e-value weighted methods a distinct power advantage over the usual normalized weights that are employed in weighted multiple testing.

Xu, Wang and Ramdas (2021) extended these results to *bandit multiple testing*. There, the data to test the K hypotheses is not available in advance, but must be collected adaptively, for example by assigning later subjects to more promising treatments as revealed by the results on earlier subjects. For each of the K treatments, one can form an e-process to test the null hypothesis that the treatment effect is nonpositive. The K e-processes have a complex dependence structure because of the adaptive assignment mechanism. Nevertheless, at any data-dependent stopping time, the e-BH procedure applied to the stopped e-processes controls the FDR.

6.3 False Coverage Rate

Suppose data regarding K parameters has been collected, a data-dependent selection rule $\mathcal{S}$ is applied to select a subset S of the parameters deemed of interest, and CIs for the selected parameters must be reported so as to keep the expected fraction of miscovering intervals at α. The BY procedure (Benjamini and Yekutieli, 2005) is an analog of the BH procedure for this task: we report $(1 - \alpha R/K)$-CIs for the selected parameters, where $1 \le R \le |S|$ is some function of the selection rule and dependence structure. Under certain dependence assumptions, this is proven to control the FCR at level α.

In contrast, the e-BY procedure of Xu, Wang and Ramdas (2022) applies only to e-CIs, which are CIs constructed by inverting tests based on e-values (discussed next). The authors prove that reporting $(1 - \alpha|S|/K)$ e-CIs controls the FCR at α for any dependence structure, and any data-dependent selection rule $\mathcal{S}$ (including one that is fully aware of the corrected intervals).

Formally, using the notation of Section 2.7, e-confidence sets for data X^τ can be defined whenever we are given a family of e-variables $\mathcal{E} := \{E^\theta_\tau : \theta \in \Theta\}$, where θ is a parameter of interest, and for all $\theta \in \Theta$, E^θ_τ is an e-variable defined relative to null $\mathbf{P}_\theta = \{P \in \Pi : \phi(\Pi) = \theta\}$. Then the $(1-\alpha)$-e-confidence set based on family $\mathcal{E}$ is simply given by $\{\theta \in \Theta : E^\theta_\tau < 1/\alpha\}$; by Markov's inequality, this is also a confidence set in the usual sense. If the e-confidence set is an interval, we refer to it as an e-CI.

Concrete examples of e-CIs include all confidence sets based on universal inference, any (arbitrarily) stopped confidence sequence, and CIs constructed using Chernoff-style concentration inequalities. Further, Xu, Wang and Ramdas (2022) show that any CI can be converted to an e-CI by calibration. A calibrator f is nonincreasing function from $[0, 1]$ to $[0, \infty)$ such that $\int_0^1 f(x)dx = 1$. If a calibrator f is also continuous at $1/\alpha$, then any CI constructed at (the more stringent) level $\alpha' := f^{-1}(1/\alpha)$ is an e-CI at level α. This e-CI is always larger than the original CI.

As before, the implications for bandit multiple testing are interesting. One can construct and continuously monitor a CS for the effect size of each treatment, decide when to stop adaptively, select any subset for further study, and report corrected CIs using the stopped CSs at the e-BY adjusted level. As one example, one could run e-BH continually using the underlying e-processes, decide when to stop based on its rejections; then the corrected CIs will be congruent with the reported discoveries in the sense that all the corrected CIs will not contain the null parameter and both FDR and FCR will be controlled at level α. (The congruence is a result of the duality between tests and CIs that we employed earlier: the e-process exceeds $1/\alpha$ if and only if the CS does not intersect the null.)

6.4 The Inevitability of e-hacking

Peeking at the data obtained so far in order to decide whether to continue is only one of many abuses of statistical testing that have been classified as “p-hacking”. Because of the interpretation in terms of betting, peeking is legitimate when we test by e-values, but other abuses are neither legitimized nor prevented. When statisticians commit these abuses using e-values, they have merely replaced “p-hacking” with “e-hacking”. A statistician is e-hacking, for example, whenever they implement many betting strategies with given data and report only the one that yields the greatest wealth.

The fundamental principle of testing by betting is that a bet on an outcome must be made before the outcome is observed. Optional continuation is allowed in the case of successive bets because this condition is still met for each individual bet. But claiming you would have bet in a certain way after you know the outcome is still humbug. We can only hope that the clarity of these principles, even for laypeople, will make the possibilities for abuse more obvious and increase the pressure to distinguish between exploratory and confirmatory analysis.

In some situations, abuses can be prevented or mitigated by a separation of roles. The use of e-values rather than p-values may be helpful in these situations.

In academic disciplines where abuses are driven by the need to publish, for example, editors can encourage preregistration of a study's data collection and analysis. If we agree that the analysis should use the betting strategy that maximizes the expected logarithm of wealth under a reasonable alternative, then the proposed analysis necessarily identifies the alleged reasonable alternative; Shafer (2021) calls this the *implied alternative*. Editors and referees could reject proposed registrations for which this implied alternative is not really plausible and even agree in advance to publish the study when it is plausible and interesting. This option does not arise when classical significance testing is used, because usually there is no unique alternative for which a test is most powerful.

When the statistician is embedded in a larger scientific enterprise, decisions about each step in data collection can be the result of consultation between the statistician and other scientists. In the first flush of excitement about Wald's sequential analysis, Barnard (1947) saw this as the future of statistics, but it has been in tension with the notion of a p-value based on a global test statistic. Testing by betting escapes this tension and can be used even in collaborative meta-analysis (Section 6.1).

7. OTHER APPLICATIONS

Game-theoretic statistics is rapidly evolving. Here are additional topics where it is relevant.

Comparing Forecasters. Many experts and pundits now repeatedly make predictions about the weather, wars, sport games, business events, and elections probabilistically, sometimes as the probability of an event (one team beating another) or a predictive distribution (over the amount of rain the next day). How can we test whether probabilistic forecasters are doing a good job (are calibrated, for example), and how can we compare two different probabilistic forecasters? Such questions have been addressed in a game-theoretic setup by several recent works that use test supermartingales (Henzi and Ziegel, 2022; Henzi, Arnold and Ziegel, 2023) or e-processes and confidence sequences (Choe and Ramdas, 2021).

A fascinating general phenomenon, called *Jeffreys's law* by Dawid (1984, Sect. 5.2) in honor of Harold Jeffreys, is that two reliable forecasters must agree in the long run: if they differ too much, a Skeptic observing both of them will be able to discredit at least one of them (Shafer and Vovk, 2019, Sect. 10.7).

Multi-Armed Bandits and Reinforcement Learning. In sequential decision making, as modeled by a contextual multi-armed bandit or a reinforcement learning problem, one sees a sequence of "contexts" $x_t \in \mathcal{X}$, and one must decide which action $a_t \in \mathcal{A}$ to take in order to maximize a (discounted) sum of observed rewards $R(x_t, a_t)$. A policy π is a mapping from $\mathcal{X}$ to $\mathcal{A}$, and one usually attempts to understand the unknown reward function R by playing some exploratory policy π_0. One central question is the following: if the data was collected using some π_0, is it possible to estimate the quality (called "value") of some other policy π_1 that was never deployed? This is called "off-policy evaluation", and is a central problem of great practical interest. Recently, Karampatziakis, Mineiro and Ramdas (2021) developed confidence sequences for off-policy evaluation when the rewards are bounded, and extended by Waudby-Smith et al. (2022) to settings with unbounded importance weights and time-varying policies. We remark that outside of the off-policy setting, CSs are very commonly designed and used in contextual bandits (Abbasi-Yadkori, Pál and Szepesvári, 2011; Chowdhury and Gopalan, 2017) and best arm identification (Jamieson et al., 2014; Kaufmann and Koolen, 2021), commonly under sub-Gaussian-type assumptions.

8. DISCUSSION

8.1 Connections with Other Areas

8.1.1 Bayesian and Evidentialist Approaches We already alluded to various connections between Bayesian and game-theoretic statistics (Bayes factors, Section 3.2.3, Jeffreys' prior, Section 3.2, Bayesian t-test Section 4.1). Even though interpretations are very different, a precise comparison would fill up an entire paper. We do highlight some relations in Appendix A, emphasizing that e-processes generalize likelihood ratios and, like these, can be interpreted as *evidence*. In the present section, we restrict ourselves to one specific simple technique to construct CSs, the *prior-posterior ratio martingale* (Waudby-Smith and Ramdas, 2020; Grünwald, 2022) so as to demonstrate how Bayesian tools can be transformed into SAVI tools. Suppose the data are drawn from P_{θ^*} for some unknown $\theta^* \in \Theta$ (extension to Bayesian nonparametrics is straightforward (Neiswanger and Ramdas, 2021)). Let $\pi_0(\cdot)$ be a "prior" distribution over Θ; we call this the working prior, because no assumptions are made about it. After seeing $X_1, X_2, \ldots, X_t$, let $\pi_t(\cdot)$ be the posterior distribution obtained via Bayes' rule. The central observation is that $\pi_0(\theta^*)/\pi_t(\theta^*)$ is a test martingale for P_{θ^*}, termed the "prior-posterior ratio martingale" (it is used for different purposes in Bayesian statistics, and is there known as the *Dickey-Savage ratio*. Thus, $\{\theta \in \Theta : \pi_0(\theta)/\pi_t(\theta) < 1/\alpha\}$ is a $(1-\alpha)$-CS for θ^*. Intriguingly, quite recently it has also been suggested within the Bayesian community Wagenmakers et al. (2020); Pawel, Ly and Wagenmakers (2022) to use this CS, when applied to fixed t, as an alternative for the Bayesian posterior credible interval; all this is further investigated in Appendix A.

8.1.2 Group-Sequential and Alpha Spending Methods. We touched on the connection to these methods in the log-rank example, Section 4.3. They are mostly used in the clinical trial literature. Like our methods, they have their roots in the work of Robbins, Siegmund, Lai and others on anytime-valid tests in the 1970s. But they developed in quite a different direction. For example—although there are exceptions[3] such as Mingxiu, Cappelleri and Gordon Lan (2007)—such methods, while providing Type-I error control under multiple looks at the data, typically require a pre-specified final sample size; whereas e-processes can continue as long as new data is available. Principles for designing e-processes, such as the GRO criteria, or the RIPr and UI methods, do not seem to have analogues in the α-spending/group sequential literature. But a firmer understanding of connections is desirable.

8.1.3 Information Theory and Online Learning. We touched on the relationship between our methods and the information-theoretic *Minimum Description Length (MDL)* paradigm for model selection, learning and prediction (Barron, Rissanen and Yu, 1998; Grünwald and Roos, 2020) when discussing the REGROW criterion in Section 3.2. The connection to MDL and the related idea of universal coding runs quite deeply, due to Kraft's inequality, which states that for any probability distribution $\bar{Q}$ with probability mass function $\bar{q}$ and any stopping time τ, there is a lossless code such for every realization x^τ, the codelength achieved with this code is equal, up to a negligible roundoff term, to $-\log \bar{q}(x^\tau)$; conversely, for any lossless code there is a distribution $\bar{Q}$ such that this correspondence holds. In MDL approaches one proceeds by associating statistical models (sets of distributions) $\mathbf{Q}$ with 'universal codes', represented as distributions $\bar{q}$ such that the codelengths are $-\log \bar{q}(x^t)$, designed to give small codelengths to the data at hand whenever the code corresponding to any element $P \in \mathbf{Q}$ assigns a small codelength to the data. This very closely mirrors the construction of $\bar{q}$ via an estimator $\hat{\theta}$ or the method of mixtures as in Section 3.2. MDL model selection between a number of parametric models $\mathbf{Q}_\gamma, \gamma \in \Gamma$ works by first associating each $\mathbf{Q}_\gamma$ with a $\bar{q}_\gamma$ as above, and then picking as 'the best explanation for data x^t' the γ for which the associated codelength $-\log \bar{q}_\gamma(x^t)$ is minimal, reporting as evidence of model $\mathbf{Q}_{\gamma_1}$ over $\mathbf{Q}_{\gamma_2}$ the codelength difference $-\log \bar{q}_{\gamma_2}(x^t) - [-\log \bar{q}_{\gamma_1}(x^t)]$. As a result, if there are just two models, $\gamma \in \{0, 1\}$ and the null

[3] We thank J. Goeman and J. ter Schure for pointing this out to us.

model is simple, the MDL approach is essentially equivalent to doing a test between null $\mathbf{Q}_0$ and alternative $\mathbf{Q}_1$ and reporting as evidence the logarithm of the e-value $\bar{q}_1(x^t)/q_0(x^t)$ (Grünwald and Roos, 2020) —exactly the same as in Section 3.2 but with evidence expressed on a logarithmic scale. When the null is composite, MDL and SAVI methods diverge, but we conjecture that e-processes have a codelength interpretation—but with different codes than in classical MDL approaches.

One may also think of 'universal codes' $\bar{q}$ as sequential prediction strategies that predict x_t using $q(x_t \mid x^{t-1})$ and with loss assessed by the *logarithmic loss function* $-\log q(x_t \mid x^{t-1})$. The vast field of *online learning* is about such sequential prediction and the logarithmic loss takes an important special case in it. Not surprisingly then, sequential prediction strategies from the online learning literature can often be converted to provide good (in some cases optimal in some sense) betting strategies for several problems. These connections have been emphasized by Orabona and Jun (2021), and also exploited by Waudby-Smith and Ramdas (2023); Shekhar and Ramdas (2021); Ramdas et al. (2022); Casgrain, Larsson and Ziegel (2022) and others.

8.2 Open Questions

8.2.1 Existence of e-processes. For what classes of distributions $\mathbf{P}$ are there nontrivial (a) test martingales, (b) test supermartingales but no test martingales, (c) e-processes but no test supermartingales, (d) none of the above? While Ramdas et al. (2022); Ruf et al. (2022) have interesting examples separating the above concepts, this separation is not yet fully understood in general.

8.2.2 Choice of Filtration. The choice of a filtration is a design choice, and one need not choose the richest one, the one generated by the observations. The choice affects both safety (the set of stopping times τ for which the expected value of E_τ does not exceed one under the null) and power (how fast the wealth grows under the alternative).

Recall the example of testing exchangeability in Section 5.5. As explained, there is no nontrivial test martingale for the problem in the filtration of the observations, but there is one in the smaller filtration of conformal p-values. This is true for any observation space, but the shrinking sacrifices safety. This sacrifice is unnecessary for small discrete alphabets, where an e-process is available in the original filtration, that appears in experiments to be at least as powerful as the conformal test martingale.

The picture is a little different for the test martingales in a shrunk filtration constructed by Pérez-Ortiz et al. (2022) for the problem of testing group-invariant hypotheses (such as the t-test example of Cox (1952) and Lai (1976) discussed in Section 4.1). These are not e-processes with respect to the original filtration, but are in the shrunk one, and thus have a weaker guarantee under the null, but they still maximize the rate of growth amongst all e-processes, even those with respect to the original filtration. Thus, they have worse safety properties but better growth properties than competitors like universal inference.

When can one shrink the filtration in order to design useful new e-processes, and when are these more or less powerful than ones in the original filtration?

8.2.3 Admissibility. Can one characterize admissibility of an e-process succinctly, with a condition that is both necessary and sufficient? Ramdas et al. (2020) define an an e-process $E \equiv (E_t)_{t\geq 1}$ for $\mathbf{P}$ to be *inadmissible* if there exists another e-process E' for $\mathbf{P}$ such that $E' \geq E$ ($E'_t \geq E_t$ almost surely P, for all $P \in \mathbf{P}$ and all $t \geq 1$), and $E'_t > E_t$ with positive probability under some $P \in \mathbf{P}$ and some $t \geq 1$; E is admissible if it is not inadmissible, meaning no such E' exists. Ramdas et al. (2020) provide both necessary and sufficient conditions for admissibility, but currently these do not match. For example, they prove that, if there exists a common dominating measure, then E being admissible implies that $E_t = \inf_{P \in \mathbf{P}} M_t^P$, where M_t^P is a test martingale for P. The universal inference e-process has this form. But this condition is not sufficient for admissibility: e-processes satisfying $E_t = \inf_{P \in \mathbf{P}} M_t^P$ may not always be admissible (indeed, universal inference has this form, and we know examples where it is inadmissible). For admissibility, the $\{M_t^P\}_{P \in \mathbf{P}}$ need to agree to some extent—they need to be large or small on similar events; if on each event, some test martingales are large while others are small, the infimum will always be small. Given that admissibility is a low bar, delineating this need for agreement is an important open problem.

8.2.4 Questions about RIPr. When exactly does the RIPr procedure applied to data X^t separately for each sample size t yield an e-process? (Section 3.3). When the RIPr yields an e-process, there is strong justification to use it, but how much is lost if it is replaced by the (always applicable) universal inference e-process? Understanding the power of universal inference is itself quite open; progress was made in the Gaussian setting by Dunn et al. (2022). In current applications the GRO-optimal RIPr e-variables are sometimes given by simple, analytic formulas (Section 4.2 and 4.1), but for other applications numerical optimization is required (Section 4.3). An algorithm presented by Li (1999) can be used for this, but it is slow. Do there exist practically effective algorithms?

REFERENCES

Abbasi-Yadkori, Y., Pál, D. and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems* **24**.

Anscombe, F. J. (1954). Fixed-sample size analysis of sequential observations. *Biometrics* **10** 89–100.

Balsubramani, A. and Ramdas, A. (2016). Sequential nonparametric testing with the law of the iterated logarithm. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*.

Barnard, G. A. (1947). Review of Abraham Wald's *Sequential Analysis*. *Journal of the American Statistical Association* **42** 658–665.

Barron, A., Rissanen, J. and Yu, B. (1998). The Minimum Description Length principle in coding and modeling. *IEEE transactions on information theory* **44** 2743-2760. Special Commemorative Issue: Information Theory: 1948-1998.

Bates, S., Jordan, M. I., Sklar, M. and Soloff, J. (2022). Principal-Agent Hypothesis Testing. *arXiv:2205.06812*.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B* **57** 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29** 1165–1188.

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association* **100** 71–81.

Berger, J. O., Pericchi, L. R. and Varshavsky, J. A. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhyā: The Indian Journal of Statistics, Series A* 307–321.

Breiman, L. (1961). Optimal gambling systems for favorable games. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*.

Carney, D. R. My position on "Power Poses". Accessed 5 June 2022, Web link.

Carney, D. R., Cuddy, A. J. C. and Yap, A. J. (2010). Power posing: Brief nonverbal displays cause changes in neuroendocrine levels and risk tolerance. *Psychological Science* **21** 1363–1368.

Casgrain, P., Larsson, M. and Ziegel, J. (2022). Anytime-valid sequential testing for elicitable functionals via supermartingales. *arXiv:2204.05680*.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques* **48** 1148–1185.

Choe, Y. J. and Ramdas, A. (2021). Comparing Sequential Forecasters. *arXiv:2110.00115*.

Chowdhury, S. R. and Gopalan, A. (2017). On kernelized multi-armed bandits. In *International Conference on Machine Learning* 844–853. PMLR.

Cox, D. R. (1952). Sequential tests for composite hypotheses. *Mathematical Proceedings of the Cambridge Philosophical Society* **48** 290–299.

Crane, H. and Shafer, G. (2020). Risk is random: The magic of the d'Alembert. Working Paper #57 at www.probabilityandfinance.com.

Darling, D. A. and Robbins, H. (1967). Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences of the United States of America* **58** 66–68.

Darling, D. and Robbins, H. (1968). Some nonparametric sequential tests with power one. *Proceedings of the National Academy of Sciences* **61** 804–809.

Dawid, A. P. (1984). Present position and potential developments: some personal views, statistical theory, the prequential approach. *Journal of the Royal Statistical Society A* **147** 278–292.

Dawid, A. P., De Rooij, S., Shafer, G., Shen, A., Vereshchagin, N. and Vovk, V. (2011). Insuring against loss of evidence in game-theoretic probability. *Statistics & Probability Letters* **81** 157–162.

De Heide, R. and Grünwald, P. D. (2021). Why optional stopping can be a problem for Bayesians. *Psychonomic Bulletin & Review* **28** 795–812.

de la Peña, V. H. (1999). A general class of exponential inequalities for martingales and ratios. *The Annals of Probability* **27** 537–564.

Delyon, B. (2009). Exponential inequalities for sums of weakly dependent variables. *Electronic Journal of Probability* **14** 752–779.

Dimitrov, V., Shafer, G. and Zhang, T. (2022). The martingale index. Working Paper #61 at www.probabilityandfinance.com.

Duan, B., Ramdas, A. and Wasserman, L. (2022). Interactive rank testing by betting. In *Proceedings First conference on Causal Learning and Reasoning*. PMLR.

Duan, B., Ramdas, A., Balakrishnan, S. and Wasserman, L. (2020). Interactive martingale tests for the global null. *Electronic Journal of Statistics* **14** 4489–4551.

Dubins, L. E. and Savage, L. J. (1965). A Tchebycheff-like inequality for stochastic processes. *Proceedings of the National Academy of Sciences* **53** 274–275.

Dunn, R., Gangrade, A., Wasserman, L. and Ramdas, A. (2021). Universal inference meets random projections: a scalable test for log-concavity. *arXiv:2111.09254*.

Dunn, R., Ramdas, A., Balakrishnan, S. and Wasserman, L. (2022). Gaussian universal likelihood ratio testing. *Biometrika*.

Efron, B. (1969). Student's t-test under symmetry conditions. *Journal of the American Statistical Association* **64** 1278–1302.

Fan, X., Grama, I. and Liu, Q. (2015). Exponential inequalities for martingales with applications. *Electronic Journal of Probability* **20** 1–22.

Feller, W. K. (1940). Statistical aspects of ESP. *The Journal of Parapsychology* **4** 271–298.

Grünwald, P. (2022). Beyond Neyman-Pearson. *arXiv:2205.00901*.

Grünwald, P., De Heide, R. and Koolen, W. (2023). Safe Testing. *Journal of the Royal Statistical Society, Series B (to appear, with discussion)*.

Grünwald, P. and Roos, T. (2020). Minimum Description Length Revisited. *International Journal of Mathematics for Industry* **11**.

Hendriks, H. (2018). Test Martingales for bounded random variables. *arXiv:1801.09418*.

Henzi, A., Arnold, S. and Ziegel, J. F. (2023). Sequentially valid tests for forecast calibration. *The Annals of Applied Statistics*.

Henzi, A. and Ziegel, J. F. (2022). Valid sequential inference on probability forecast performance. *Biometrika*.

Howard, S. R. and Ramdas, A. (2022). Sequential estimation of quantiles with applications to A/B testing and best-arm identification. *Bernoulli* **28** 1704–1728.

Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2020). Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys* **17** 257–317.

Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics* **49** 1055–1080.

Ignatiadis, N., Wang, R. and Ramdas, A. (2022). E-values as unnormalized weights in multiple testing. *arXiv:2204.12447*.

Jamieson, K., Malloy, M., Nowak, R. and Bubeck, S. (2014). lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory* 423–439. PMLR.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Oxford University Press, London.

Johari, R., Koomen, P., Pekelis, L. and Walsh, D. (2022). Always valid inference: Continuous monitoring of A/B tests. *Operations Research* **70** 1806–1821.

John, L. K., Loewenstein, G. and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science* **23** 524–532.

Karampatziakis, N., Mineiro, P. and Ramdas, A. (2021). Off-policy confidence sequences. In *International Conference on Machine Learning* 5301–5310. PMLR.

Kaufmann, E. and Koolen, W. M. (2021). Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals. *Journal of Machine Learning Research* **22** 246–1.

Kelly, J. L. (1956). A New Interpretation of Information Rate. *Bell System Technical Journal* 917–926.

Lai, T. L. (1976). On confidence sequences. *The Annals of Statistics* **4** 265–280.

Lhéritier, A. and Cazals, F. (2018). A sequential non-parametric multivariate two-sample test. *IEEE Transactions on Information Theory* **64** 3361–3370.

Li, J. Q. (1999). Estimation of Mixture Models, PhD thesis, Yale University, New Haven, CT.

Li, J. Q. and Barron, A. R. (2000). Mixture Density Estimation. In *Advances in Neural Information Processing Systems* **12** 279–285.

Lindon, M. and Malek, A. (2020). Anytime-Valid Inference for Multinomial Count Data. *arXiv:2011.03567*.

Manole, T. and Ramdas, A. (2023). Sequential estimation of convex divergences using reverse submartingales and exchangeable filtrations. *IEEE Transactions on Information Theory*.

Mingxiu, H., Cappelleri, J. C. and Gordon Lan, K. K. (2007). Applying the law of iterated logarithm to control type I error in cumulative meta-analysis of binary outcomes. *Clinical Trials*.

Neiswanger, W. and Ramdas, A. (2021). Uncertainty quantification using martingales for misspecified Gaussian processes. In *Algorithmic Learning Theory* 963–982. PMLR.

Orabona, F. and Jun, K.-S. (2021). Tight concentrations and confidence sequences from the regret of universal portfolio. *arXiv:2110.14099*.

Pace, L. and Salvan, A. (2020). Likelihood, Replicability and Robbins' Confidence Sequences. *International Statistical Review* **88** 599–615.

Pandeva, T., Bakker, T., Naesseth, C. A. and Forré, P. (2022). E-Valuating Classifier Two-Sample Tests.

Pawel, S., Ly, A. and Wagenmakers, E.-J. (2022). Evidential Calibration of Confidence Intervals. *arXiv:2206.12290*.

Pérez-Ortiz, M. F., Lardy, T., De Heide, R. and Grünwald, P. (2022). E-Statistics, Group Invariance and Anytime Valid Testing. *arXiv:2208.07610*.

Podkopaev, A., Bloebaum, P., Kasiviswanathan, S. and Ramdas, A. (2023). Sequential kernelized independence testing. In *International Conference on Machine Learning*.

Ramdas, A., Ruf, J., Larsson, M. and Koolen, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv:2009.03167*.

Ramdas, A., Ruf, J., Larsson, M. and Koolen, W. M. (2022). Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning* **141** 83–109.

Rissanen, J. (1984). Universal coding, information, prediction and estimation. *IEEE Transactions on Information Theory* **30** 629–636.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* **58** 527–535.

Robbins, H. and Siegmund, D. (1974). The expected sample size of some tests of power one. *The Annals of Statistics* **2** 415–436.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. and Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review* **16** 225–237.

Royall, R. (1997). *Statistical evidence: a likelihood paradigm*. Chapman and Hall.

Ruf, J., Larsson, M., Koolen, W. M. and Ramdas, A. (2022). A composite generalization of Ville's martingale theorem. *arXiv:2203.04485*.

Rushton, S. (1950). On a Sequential t-test. *Biometrika* **37** 326–333.

Shafer, G. (2021). Testing by betting: a strategy for statistical and scientific communication (with discussion and response). *Journal of the Royal Statistic Society A* **184** 407–478.

Shafer, G. and Vovk, V. (2001). *Probability and Finance: It's Only a Game*. Wiley, New York.

Shafer, G. and Vovk, V. (2019). *Game-Theoretic Foundations for Probability and Finance*. Wiley, Hoboken, New Jersey.

Shafer, G., Shen, A., Vereshchagin, N. and Vovk, V. (2011). Test martingales, Bayes factors and p-values. *Statistical Science* **26** 84–101.

Shekhar, S. and Ramdas, A. (2021). Nonparametric two sample testing by betting. *arXiv:2112.09162*.

Shin, J., Ramdas, A. and Rinaldo, A. (2022). E-detectors: a nonparametric framework for online changepoint detection. *arXiv:2203.03532*.

Spertus, J. V. and Stark, P. B. (2022). Sweeter than SUITE: Supermartingale Stratified Union-Intersection Tests of Elections. In *International Joint Conference on Electronic Voting*.

Ter Schure, J. and Grünwald, P. (2022). ALL-IN meta-analysis: breathing life into living systematic reviews. *F1000Research* **11**.

Ter Schure, J., Perez-Ortiz, M. F., Ly, A. and Grünwald, P. (2021). The Safe Log Rank Test: Error Control under Continuous Monitoring with Unlimited Horizon. *arXiv:1906.07801*.

Turing, A. M. (c 1941). The Applications of Probability to Cryptography. UK National Archives, HW 25/37. See arXiv:1505.04714 for a version set in Latex.

Turner, R. and Grünwald, P. (2023). Anytime-valid Confidence Intervals for Contingency Tables and Beyond. *Statistics and Probability Letters*.

Turner, R., Ly, A. and Grünwald, P. (2021). Generic E-Variables for Exact Sequential k-Sample Tests that allow for Optional Stopping. *arXiv:2106.02693*.

Turner, R., Ly, A., Ortiz-Perez, M.-F., ter Schure, J. and Grünwald, P. (2022). R-package `safestats`. CRAN.

Ville, J. (1939). *Etude critique de la notion de collectif*. Gauthier-Villars.

Volkhonskiy, D., Burnaev, E., Nouretdinov, I., Gammerman, A. and Vovk, V. (2017). Inductive conformal martingales for change-point detection. In *Conformal and Probabilistic Prediction and Applications* 132–153. PMLR.

Vovk, V. (2021). Testing randomness online. *Statistical Science* **36** 595–611.

Vovk, V., Nouretdinov, I. and Gammerman, A. (2021). Conformal testing: binary case with Markov alternatives. *arXiv:2111.01885*.

Vovk, V. and Wang, R. (2021). E-values: Calibration, combination, and applications. *The Annals of Statistics* **49** 1736–1754.

Wagenmakers, E.-J., Gronau, Q. F., Dablander, F. and Etz, A. (2020). The support interval. *Erkenntnis* 1–13.

Wald, A. (1945). Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics* **16** 117-186.

Wald, A. (1947). *Sequential Analysis*. Wiley, New York.

Wang, R. and Ramdas, A. (2022). False discovery rate control with e-values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Wang, H. and Ramdas, A. (2023). Catoni-style confidence sequences for heavy-tailed mean estimation. *Stochastic Processes and Applications*.

Wasserman, L., Ramdas, A. and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences* **117** 16880–16890.

Waudby-Smith, I. and Ramdas, A. (2020). Confidence sequences for sampling without replacement. In *Advances in Neural Information Processing Systems* **33** 20204–20214.

Waudby-Smith, I. and Ramdas, A. (2023). Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. (to appear with discussion).

Waudby-Smith, I., Stark, P. B. and Ramdas, A. (2021). RiLACS: Risk limiting audits via confidence sequences. In *International Joint Conference on Electronic Voting* 124–139. Springer.

Waudby-Smith, I., Arbour, D., Sinha, R., Kennedy, E. H. and Ramdas, A. (2021). Time-uniform central limit theory and asymptotic confidence sequences. *arXiv:2103.06476*.

Waudby-Smith, I., Wu, L., Ramdas, A., Karampatziakis, N. and Mineiro, P. (2022). Anytime-valid off-policy inference for contextual bandits. *arXiv preprint arXiv:2210.10768*.

Xu, Z., Wang, R. and Ramdas, A. (2021). A unified framework for bandit multiple testing. In *Advances in Neural Information Processing Systems* **34**.

Xu, Z., Wang, R. and Ramdas, A. (2022). Post-selection inference for e-value based confidence intervals. *arXiv:2203.12572*.

APPENDIX A: SAVI AS A FREQUENTIST – EVIDENTIAL – BAYESIAN MIDDLE GROUND?

E-values can be seen as a evidence against a null hypothesis and are quite meaningful even without being used for a sequential test required to have some error probability, and even in a batch setting such as the multiple testing settings above. E-values lack some of the properties of p-values that make the latter less suitable to think of as 'evidence' (such as the p-value's dependency on whether or not particular actions are taken in counterfactual situations) and generalize the likelihood ratio that is embraced by the likelihoodists as the 'right' formalization of relative evidence (Royall, 1997).

Comparing our methods to Bayesian ones, we see that, with simple nulls, e-processes and Bayes factors always coincide; in parametric tests with composite nulls, e-processes and Bayes factors sometimes (e.g in the group invariant setting of Section 4.1) but not always coincide; and with nonparametric tests they start differing quite a lot. If it comes to confidence sequences and e-confidence intervals, we find that even in one-dimensonal parametric settings, $(1-\alpha)$- e-confidence intervals (and equivalently stopped confidence sequences) do not coincide with Bayesian $(1-\alpha)$-posterior credible intervals, the latter being significantly narrower. To see this, note that, from Section 3.2 (and defining e-CIs as in Section 6.3), we find that, for any fixed prior density w on $\Theta \subset \mathbb{R}$, the family of e-variables $\{E^{\theta}_{\tau} : \theta \in \Theta\}$ for data X^{τ} with

$$\bar{P}(\theta \mid X^{\tau}) := \frac{1}{E^{\theta}_{\tau}} = \frac{p_{\theta}(X^{\tau})}{\int p_{\theta'}(X^{\tau}) w(\theta') d\theta'}$$

(this is just the reciprocal of the prior-posterior ratio martingale of Section 8.1.1 stopped at time τ) defines an e-confidence interval at level $(1-\alpha)$ as $\{\theta : \bar{P}(\theta \mid \tau) \geq \alpha\}$, whenever the latter set is an interval. By Bayes'theorem, the Bayes posterior based on the same prior w is given by

$$w(\theta \mid X^{\tau}) = \frac{w(\theta) \cdot p_{\theta}(X^{\tau})}{\int p_{\theta'}(X^{\tau}) w(\theta') d\theta'} = w(\theta) \cdot \bar{P}(\theta \mid X^{\tau}),$$

and defines a posterior credible interval at level $(1-\alpha)$ as $[\theta_L, \theta_R]$ chosen so that

$$\mathbb{E}_{\theta \sim W}[\mathbf{1}_{\theta \in [\theta_L, \theta_R]} \cdot \bar{P}(\theta \mid X^{\tau})] = \int_{\theta_L}^{\theta_R} w(\theta \mid X^{\tau}) = 1 - \alpha.$$

We see that all elements θ of an e-confidence interval must have $\bar{P}(\theta \mid X^{\tau}) \geq \alpha$; for a Bayesian credible interval this only has to hold in average over the prior, causing the latter to be narrower in practice.

Intriguingly though, some Bayesian statisticians have noted that the standard Bayesian posterior credible interval has no clear 'evidential' interpretation. They instead propose a *Bayesian support interval* (Wagenmakers et al., 2020)—also called 'evidential support interval'— where the k- support interval is the interval containing all parameter values under which the observed data X^{τ} are at least k times as likely than under the Bayesian marginal distribution'. As Pawel, Ly and Wagenmakers (2022) note, for simple nulls and $k < 1$, this actually coincides precisely with the $(1-k)$-e-confidence interval based on the family of e-values based on the same prior density w as the Bayes marginal. If one were to consider growing sequences of data and use the same prior at each sample size t, the resulting process of support intervals would also be a standard confidence sequence in our sense.

We would venture that, for models Π with parameter of interest $\theta = \phi(P)$ and additional nuisance parameters, and also in nonparametric settings, the e-confidence intervals based on a family $\{E^{\theta}_{\tau} : \theta \in \Theta\}$ still have an 'evidential' interpretation, although in this case they will usually not be equal to a Bayesian support interval any more.

Taking the 'e-values are similar to, but different from Bayes factors' line of reasoning even further, one could daringly suggest to define $\bar{P}(\theta \mid X^{\tau}) := 1/E^{\theta}_{\tau}$ as an analogue of the Bayesian posterior or confidence distributions, even for multiparameter and nonparametric problems in which it does not coincide with the Savage-Dickey density ratio. This was done informally in (Waudby-Smith and Ramdas, 2020, Appendix E7) who visualize uncertainty by drawing $\bar{P}(\theta \mid X^{\tau})$ as a function of θ. Grünwald (2022) shows that this *e-posterior* can be motivated not just evidentially, but also decision-theoretically. Just like the Bayes posterior can, assuming the prior was chosen well, be used to obtain optimal decisions for arbitrary loss functions by combining posterior and loss in a certain way (minimizing Bayes-posterior expected loss), the e-posterior can be used as a basis for obtaining decisions with minimax optimality guarantees for arbitrary loss functions by combining e-posterior and loss in a certain, different way. The guarantees hold irrespective of the chosen prior, but become weaker the more atypical the data look with respect to the prior. In the same paper (see also Bates et al. (2022) for related insights), it is shown that, even in a nonsequential context, standard Neyman-Pearson testing is not adequate if the decision problem at hand (e.g. choose between the four actions { vaccinate noone; only adults; only the elderly; or everyone}) has more actions than just the 'reject' and 'accept' of the Neyman-Pearson theory; with a decision rule based on e-variables one can effectively deal with such—realistic—settings. This has direct repercussions for the reproducibility crisis: the standard Neyman-Pearson based approaches may simply not be suitable for the complex real-world problems that we apply our test results to.

In conclusion, let us stress that we do not view the above observations as disqualifying the Bayesian, evidential or Neyman-Pearsonian paradigm. Rather, we feel that SAVI methods effectively unify some of the fundamental ideas of each; respectively: one should allow for the possibility to infuse prior knowledge into one's procedures; one should output numbers with a clear evidential meaning; and one should ensure that one's procedures allow for error control and coverage.