Statistical Science
2023, Vol. 38, No. 4, 576–601
https://doi.org/10.1214/23-STS894

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 crucially rely 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 centrally employs game-theoretic intuition, language and mathematics. We summarize the SAVI goals and philosophy, and report recent advances in testing composite hypotheses and estimating functionals in nonparametric settings.

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. (2021a), 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

Aaditya Ramdas is Assistant Professor, Statistics and Data Science, and Machine Learning, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA (e-mail: aramdas@cmu.edu). Peter Grünwald is Head, Machine Learning Group, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands, and Professor of Statistics at Leiden University, The Netherlands (e-mail: pdg@cwi.nl). Vladimir Vovk is Professor, Computer Science, Royal Holloway, University of London, UK (e-mail: v.vovk@rhul.ac.uk). Glenn Shafer is University Professor, Rutgers University, Piscataway, New Jersey 08854-8019, USA (e-mail: gshafer@business.rutgers.edu).

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.

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. E-processes permit 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 Appendix B.

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 many 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 settings that have emerged with the development of information technology in the past half-century, including "living meta-analysis" (Ter Schure, Grünwald and Ly, 2021), the industrial use of A/B testing (Johari et al., 2022) and bandit experiments (Howard and Ramdas, 2022).

The new methods can be most clearly presented in the language of game-theoretic probability (Shafer and Vovkm 2001a, 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 under forecaster's probabilities) 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 (2001a, 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.

1.1 Basic Terminology

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 (i.i.d.) under P. We use X^t as a shorthand for $X_1, \dots, X_t$.

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 nonintersecting or nested subsets of Π. We always use

boldface $\mathbf{P}$, $\mathbf{Q}$ for sets of distributions, and normal P, Q for a single distribution.

A sequence of random variables $Y \equiv (Y_t)_{t\geq 0}$ is called a *process* if it is adapted to $\mathbf{F}$—that is, 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 ($\mathbf{F} = \{\varnothing, \Omega\}$), and in this case Y_t being measurable with respect to $\mathbf{F}_t$ means that Y_t is a measurable function of $X_1, \ldots, X_t$. But $\mathbf{F}$ is sometimes a coarser filtration (we discard information, see, for example, Section 4.1.2) or a richer one (we add external randomization).[1] Y is called *predictable* if Y_t is measurable with respect to $\mathbf{F}_{t-1}$.

A stopping time (or rule) τ is a nonnegative integer valued random variable such that $\{\tau \leq t\} \in \mathbf{F}_t$ for each $t \geq 0$. 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.

1.2 A Terse Technical Summary of the Paper

We give a short technical summary below, foreshadowing topics to be defined and discussed in more depth later.

The field of safe anytime-valid inference (SAVI) aims to develop measures of statistical evidence and certainty that remain valid at arbitrary stopping times (possibly unknown in advance), accommodating continuous monitoring and analysis of accumulating data and optional stopping or continuation for any reason. There is a strong sense in which *admissible* SAVI methods—power-one tests, confidence sequences, anytime-valid p-values and e-processes—*must* rely centrally on nonnegative martingales (Ramdas et al., 2020). Nonnegative (super)martingales are endowed with a strong and direct connection to gambling: every nonnegative supermartingale corresponds to a wealth process in some game, and vice versa (every "fair/legal" gambling strategy to test the null hypothesis results in a wealth process that is a nonnegative supermartingale).

These facts give rise to the following central principle in game-theoretic statistics: "testing by betting". In order to test a null hypothesis $\mathbf{P}$ against an alternative $\mathbf{Q}$, we set up a game such that (a) if the null is true, meaning $P \in \mathbf{P}$, then no betting strategy can reliably make money (any gambler's wealth is a nonnegative supermartingale), and (b) if the null is false, meaning $P \in \mathbf{Q}$, it is possible to bet smartly to make money in that game. This principle arguably has roots dating back (at least) to Ville (1939), and was recently discussed in depth in the point null case by Shafer et al. (2011) and Shafer (2021) and for composite nulls by Grünwald, De Heide and Koolen (2023), Waudby-Smith and Ramdas (2023), etc.

[1]The filtration may be coarsened, as explained by Alan Turing (1941, page 1): "When the whole evidence…is taken into account it may be extremely difficult to estimate the probability of the event, …may be better to form an estimate based on a part of the evidence …".

The game works as follows. Before observing $X_t \in \mathcal{X}$, Skeptic puts forward a bet $S_t : \mathcal{X} \to [0, \infty]$, which satisfies

$$\mathbb{E}_P\big[S_t(X_t)\big] \leq 1 \quad \text{for every } P \in \mathbf{P}. \tag{1}$$

Then, X_t is revealed. The interpretation is that at each time t, one can buy, at the price of 1 monetary unit, a ticket that will pay off $S_t(X_t)$ units. One can buy as many tickets as one likes. (1) simply expresses that under the null, one does not expect to get back more than one invests in this game. At time 1, Skeptic invests 1 monetary unit; at each time t, she reinvests all the money she observed so far. Skeptic's wealth after t steps is then clearly given by $\prod_{i=1}^{t} S_i(X_i)$, which is nonnegative by definition, and easily checked to be a supermartingale under P. If $\mathbf{Q}$ is appropriately separated from $\mathbf{P}$, good betting strategies can force the wealth in setting (b) (alternative is true) to grow to infinity exponentially fast, and we wish to maximize the exponent. Maximizing the exponent corresponds to maximizing the expected logarithm of the wealth; such a "log-optimality" objective has information-theoretic roots dating back to Kelly (1956) and Breiman (1961), but also (implicitly or explicitly) appears in the work of Ville, Wald, Robbins, etc.

For testing a point null P against a point alternative Q, the log-optimal bet is simply given by the likelihood ratio $S_t = dQ_t/dP_t$, where P_t and Q_t are the conditional probabilities for the tth observation given the past, under P and Q respectively. Thus, the realized likelihood ratio of Q against P is precisely the optimal wealth of a gambler betting against P. This central fact provides much intuition for extensions and generalizations.

For composite alternatives $\mathbf{Q}$, the Skeptic often hedges their bets by not betting all their money on a single $Q \in \mathbf{Q}$, instead spreading their investment over $\mathbf{Q}$ using a mixture ("prior") distribution R. To illustrate using the toy case in which $\mathbf{Q}$ is countable and R has mass function r, $r(q) = a$ would mean that a fraction a of Skeptic's money is invested in q—importantly in general R neither has a frequentist ('drawing from an urn') nor a Bayesian ('belief') interpretation here. This *method of mixtures* plays a central role in this paper: an instance of Laplace's method for approximating a maximum by an integral, it appears directly within our anytime-valid context in Ville (1939) and Robbins (1970), and in broader sequential contexts in Wald (1947) and Cover (1974), among many others.

The most interesting questions in this area involve composite (and often nonparametrically specified) nulls $\mathbf{P}$. Indeed, there really was no general theory for dealing with composite nulls until 2017—when, almost out of the blue, several generic proposals for dealing with composite nulls appeared. Arguably, it is this development which caused the aforementioned explosion of interest in the area—suddenly there was an indication that eventually almost

any interesting statistical testing or estimation problem could be converted into an anytime-valid version with a gambling interpretation.

For such composite **P**, a fascinating phenomenon sometimes presents itself: for some "extremely rich" nulls **P**, the game described above is hopelessly restraining: the constraint (1) is too stringent, and the only functions S_t that satisfy it are either constant or decreasing (meaning that they cannot increase under any alternative). This happens, for example, when testing exchangeability or testing log-concavity; see Section 5.5 for references and details.

Luckily, generalizing the above game protocol resuscitates the approach. There appear to be two different types of generalized games: (a) one can restrict the amount of information available to the Skeptic by introducing a third player (an "Intermediary") who throws away some information revealed by Reality (mathematically, Skeptic operates in a shrunk filtration), (b) one can instead make the Skeptic play many games in parallel, each against a different subset of **P**, with the Skeptic's net wealth being their worst wealth across all the parallel games. In the first case, Skeptic's wealth may remain a nonnegative (super)martingale, but in the second case, their wealth is an *e-process* (under the null, their wealth is upper-bounded by a different supermartingale in each game, and thus is bounded by one at any stopping time). While these solutions may seem almost magical at first glance, they both yield fruit for the same problem mentioned above of testing exchangeability: approach (a) is used in Vovk, Gammerman and Shafer (2022, Part III) and approach (b) in Ramdas et al. (2022). The latter work, along with Ruf et al. (2022), together show the centrality of e-processes in game-theoretic statistics: e-processes exist for many **P** for which nonnegative (super)martingales do not.

When **P** and **Q** have a common reference measure, meaning that likelihood-ratio based methods are still in play, two key ideas stand out: universal inference (Wasserman, Ramdas and Balakrishnan, 2020), and the reverse information projection (Grünwald, De Heide and Koolen, 2023). The former always yields an e-process, but latter always results in an e-value which can be multiplied across batches of data to yield a supermartingale. But *sometimes* the latter also directly yields an e-process (and when it does, it dominates universal inference).

When **P** and **Q** do not have any common reference measures—and thus likelihood-ratio based methods may not make any sense at the outset—the *design* of nonnegative (super)martingales or e-processes occupies center stage. Sometimes, the nonparametric definition of **P** directly yields a natural game, like when testing if a "sub-Gaussian" mean is positive (Darling and Robbins, 1967). Other times, one must design new games in possibly shrunk filtrations, which may not be obvious at the outset, like in two-sample testing (Shekhar and Ramdas, 2023a).

The entire discussion above was centered on testing by betting, because this typically forms the technical heart of other problems that are not cast explicitly as testing. For example, appropriate duality concepts and inversions allow us to translate many of these results into those for estimation of appropriate functionals using confidence sequences (Section 5). Both e-processes and confidence sequences can in turn be extended to other problems like change detection (Section 5.7), model selection, etc.

In fact, our investigations reveal a curious phenomenon: at the *heart* of many (and plausibly, *all*) nonparametric testing and estimation problems is a "hidden" game (often not unique: the same **P** and **Q** may be associated with different filtrations and betting strategies that are e-processses under **P** and make money under **Q**). Further, explicitly bringing out such games (and betting well in them) can yield powerful new methodology as well as new theoretical insights (Howard et al., 2020)

A full understanding of when and why this happens is open, but we provide one hint here. Likelihood ratios have been at the center of statistics for nearly a century. Nonnegative (super)martingales and e-processes are simply nonparametric, composite generalizations of likelihood ratios, and these have been found to exist in dozens of problems where one cannot even begin to talk about likelihood-ratio based methods. Thus, these tools give us a way to work implicitly with likelihood ratios, even when there appears to be no explicit way to do so. Given the power (and sometimes optimality) of likelihood-ratio based tests in parametric settings, we perhaps get a hint of the power of our game-theoretic approaches in composite (often nonparametric) settings.

The rest of this paper will formally define the key concepts, and provide technical details of the aforementioned methods and phenomena in different problem settings.

2. CENTRAL CONCEPTS

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 **P** is a nonnegative random variable E such that $\mathbb{E}_P[E] \leq 1$ for all $P \in \mathbf{P}$. Its realized value, after observing the data, is an *e-value*.[2] 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 often used for both 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

[2]Observe that we use boldface $\mathbb{E}$ for expectation and normal E for e-values. The "e" in e-value stands both for "evidence" (because it quantifies statistical evidence against the null) and for "expectation" (because its central property is its expectation).

values are prices of payoff: the Forecaster predicts that $X \sim P$, and in order to bet against them, a Skeptic could buy one unit of E, for the price of 1, delivering the Skeptic a payoff of $E(X)$. Mathematically, a unit bet against P is simply[3] 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 $\int E(x)p(x)\,dx = 1$, 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 $\mathbb{E}_P[q/p] = \int p(x)\frac{q(x)}{p(x)}\,dx = 1$.

We use e-values when data are treated as a batch. Their dynamic counterparts are test martingales and e-processes, introduced next.

2.2 Test (Super)Martingales

A process M is a *martingale* for P if

$$\mathbb{E}_P[M_t|\mathbf{F}_{t-1}] = M_{t-1} \tag{2}$$

for all $t \geq 1$. M is a *supermartingale* for P if it satisfies (2) with "=" relaxed to "$\leq$". 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 $t \geq 0$, 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—$\mathbb{E}_P[M_\tau] \leq 1$, 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

$$\frac{q(X^t)}{p(X^t)} = \frac{q(X_1)}{p(X_1)}\frac{q(X_2|X_1)}{p(X_2|X_1)}\cdots\frac{q(X_t|X^{t-1})}{p(X_t|X^{t-1})}, \tag{3}$$

where $X^t := (X_1, \ldots, X_t)$, $p(X^t)$ is P's density for X^t, and $q(X^t)$ is Q's density for X^t. Denote the sequence defined by (3) as M; then M is a test martingale for P, and

$$M_t = \prod_{i=1}^{t} B_i = \frac{q(X^t)}{p(X^t)}, \tag{4}$$

$$\text{where} \quad B_t := \frac{q(X_t|X^{t-1})}{p(X_t|X^{t-1})}. \tag{5}$$

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*.

[3]In some sense, statisticians have always been using e-values (and test martingales), because likelihood ratios are the most important example of e-values (and test martingales). But this direct analog only holds when testing a single distribution P. The power and utility of e-values, test (super)martingales and e-processes are truly realized only dealing with a composite (and sometimes nonparametric) $\mathbf{P}$.

Test martingales for P are always of the form (4). 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 (5); see Section 3.2. Test supermartingales can also be decomposed in the style of (4), where the B_t are *single-round* e-values (i.e., defined as function on a single outcome X_t) conditional on $\mathbf{F}_{t-1}$.

Test martingales become more interesting objects in the composite setting.

2.3 Composite Test (Super)Martingales

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 also decompose as in (4): for every $P \in \mathbf{P}$, there is a Q that is absolutely continuous with respect to P and satisfies $M_t = q(X^t)/p(X^t)$; see Ramdas et al. (2020, Proposition 4). In other words, *composite test martingales are simultaneous likelihood ratios*.

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. In particular, we would like test martingales for $\mathbf{P}$ that increase to infinity under the alternative $\mathbf{Q}$. But 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 even these may not exist (Section 5.5). For this reason, we also need e-processes.

2.4 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

$$E_t \leq M_t^P \quad \text{for every } P \in \mathbf{P}, t \geq 0. \tag{6}$$

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

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

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*[4] e-processes (see Section 8.2.3) must satisfy

$$E_t = \inf_{P\in\mathbf{P}} M_t^P \tag{7}$$

[4]An e-process $E \equiv (E_t)_{t\geq 1}$ for $\mathbf{P}$ is *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.

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 $X_1, X_2, \ldots$ (Ramdas et al., 2022, Section 5.4).

The evidence against a null hypothesis as measured by a test martingale or e-process may decrease as we collect more data (indeed, gamblers may lose money as they play a game, even if the odds are stacked in their favor). In order to obtain a measure of evidence that does not decrease with time, one can calculate the running maximum of the e-process $(\sup_{s\le t} E_s)$—which is not an e-process—and then adjust it back to being an e-process using a lookback calibrator; see Shafer et al. (2011), Dawid et al. (2011) and Ramdas et al. (2022, Section 4.7).

2.5 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 probability distribution 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—that is, 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.[5] These results have been called *Ville's theorem* (Shafer and Vovk (2019, Section 9.1). Ruf et al. (2022) generalize Ville's theorem to composite $\mathbf{P}$, but this cannot be accomplished by a test martingale, requiring e-processes instead.

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

$$P\Big(\sup_t M_t \ge \alpha\Big) \le \frac{1}{\alpha}. \tag{8}$$

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 α 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 holds with equality for continuous-path (hence continuous-time) test martingales with unbounded total variation: a classic example would be the process $M_t = \exp(\lambda B_t - t\lambda^2/2)$, where B_t is a standard Brownian motion and λ is any nonzero constant. In discrete time, the main source of looseness is from "overshoot": continuous path martingales equal $1/\alpha$ at the instant of crossing it (they do not overshoot), but discrete time (and thus discrete path) test martingales are typically strictly larger than $1/\alpha$ at the first time of crossing; this causes some looseness. In practice (which is always in discrete time), Ville's inequality is "quite tight" and overshoot is often considered a second order effect.

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

$$\sup_{P\in\mathbf{P}} P\big(\exists t \ge 1 : E_t \ge 1/\alpha\big) \le \alpha. \tag{9}$$

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

$$P\big(E_\tau \ge 1/\alpha\big) \le \alpha \quad \text{for every } \tau \in \mathcal{T},\ P \in \mathbf{P}. \tag{10}$$

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

Recently, Wang and Ramdas (2023a) extended Ville's inequality to apply to *nonintegrable* nonnegative supermartingales, such as those obtained when mixing likelihood ratios with an improper prior.

2.6 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 $(0, 0, 0, 0, 1, 1, 1, 1, \ldots)$, 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

$$P(\exists t \ge 0 : \psi_t = 1) \le \alpha \quad \text{for all } P \in \mathbf{P}. \tag{11}$$

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

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

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*.

Note that our sequential tests are different from Wald's original sequential tests, which dominated the area for a long time—in these, the null hypothesis may finally be accepted and *the stopping rule is specified beforehand*, determined by the desired Type-I/II error bounds α and β. Our framework resembles the "power-one tests" of Darling and Robbins (1968), where we do not commit to a stopping rule, and could keep going if we do not reject $\mathbf{P}$.

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

2.7 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. (2021a)) 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 (Howard et al., 2021a, Lemma 3), $P(\inf_t p_t \le u) = P(\exists t \ge 1 : p_t \le u) \le u$. 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.

2.8 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

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

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

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

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

$$C_t := \big\{\theta \in \Theta : M_t^\theta < 1/\alpha\big\}. \tag{12}$$

2.9 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 (e.g., 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 data; otherwise we are martingaling. Recently, Wasserman, Ramdas and Balakrishnan (2020) used such averaging techniques to derandomize universal inference (discussed in Section 3.3). 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 (1970), Robbins and Siegmund (1974)); see Howard et al. (2021a) for recent advances.

2.10 Multiplying e-Values

Independent e-values can be combined by multiplication. If $B_1, \dots, B_n$ 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 $B_1, \dots, B_{k-1}$, that is, if

$$\mathbb{E}[B_k | B_1, \dots, B_{k-1}] \le 1$$

for $k \ge 1$, then $M_n = \prod_{k=1}^n B_k$ is an e-value for $\mathbf{P}$. The sequence $(M_n)_{n\ge0}$ 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,

$$\mathbb{E}_P[S_t | \mathbf{F}_{t-1}] \le 1 \quad \text{for all } P \in \mathbf{P}. \tag{13}$$

If (13) holds for all t, then $M_n = \prod_{t=1}^n S_t$ is an e-value for $\mathbf{P}$. The sequence $(M_n)_{n\ge0}$ 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, page 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

$$\mathbb{E}_Q(\log E_t) \tag{14}$$

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 i.i.d. 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 i.i.d. 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"[6] 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. (For these reasons, the former quantity has been recently termed "e-power".)

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 (4) has a reasonably high expected growth rate under any distribution in $\mathbf{Q}$ that fits the data $X_1, X_2, \ldots$ reasonably well. Because $X_1, X_2, \ldots$ 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 (5). 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 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, equation (10.10)); it was subsequently analyzed by Robbins and Siegmund (1974), who

[6] Advocates of Kelly betting in the stock market also use half Kelly and fractional Kelly strategies, which make sense when you are not confident about the alternative Q (Maclean, Thorp and Ziemba, 2010).

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 (5) 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

$$\inf_{Q\in\mathbf{Q}} \mathbb{E}_Q\big(\log M_\tau - \gamma \cdot \log M_\tau^Q\big), \tag{15}$$

where $\gamma = 1$ and M^Q is the GRO e-process so that $M_\tau^Q = (dP/dQ)(X^\tau)$, maximizing (14) with t replaced by τ. 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*). A variation of this criterion, GROW, is obtained if we set $\gamma = 0$. We then search for the R that maximizes worst-case expected logarithmic return $\mathbb{E}_Q(\log M_\tau)$ in an absolute rather than relative sense. In most applications, we are interested in REGROW rather than GROW optimality so we will focus on it below. Still, as will become clear in Section 4, in some applications GROW is more appropriate and in some applications they in fact coincide; (Grünwald, De Heide and Koolen, 2023) discuss the differences in detail.

Under slight regularity conditions, the e-variable M_τ that maximizes (15) can be written as a Bayes factor defined relative to a specific prior R_τ, where R_τ varies with τ and, in many cases, with γ. 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 (15) for all τ. We will generalize (15) 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 4.1.2.

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 (15) 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 *When precise alternatives are hard to come by.* The previous development is based on the premise that we take the alternative $\mathbf{Q}$ very seriously, as containing distributions from which the data may actually be sampled. More generally, we may simply have a set of distributions available which we think may predict the data reasonably well in terms of the log score, but which we would never believe to be 'true', so that taking an expectation over them as in the GRO definition is not too meaningful. There is nothing that stops us from using such $\mathbf{Q}$ in combination with, for example, the plug-in method; we will still have an e-process, we merely cannot claim its

optimality any more. We return to this more 'Fisherian' perspective on testing in Section 8.1.3.

Alternatively, Waudby-Smith and Ramdas's (2023) GRAPA (Growth Rate Adaptive to the Particular Alternative) method uses, on round t, the empirical distribution of X^{t-1}, possibly smoothed, for the numerator in (5). 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

$$E_t^{\mathrm{UI}} := \frac{\bar{q}(X^t)}{\sup_{p\in\mathbf{P}} p(X^t)} = \frac{\bar{q}(X^t)}{\hat{p}_{X^t}(X^t)}, \tag{16}$$

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 $\bar{q}(X^t) := \int \prod_{i=1}^t q(X_i|X^{i-1})\, dR(q)$, 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 (6). 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) (original arXiv version 2019) 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

$$D(\bar{Q}^\tau \| P^\tau) = \inf_{W\in\mathbf{W}} D(\bar{Q}^\tau \| P_W^\tau), \tag{17}$$

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

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

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 (14) 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, 4.1.2) it is easy to calculate, in others (e.g., Cox regression (Ter Schure et al., 2021)) it is not. In general, it is a sub-probability measure, that is, 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 of the set of mixtures of $\mathbf{P}$, the infimum is often achieved by some $W \in \mathbf{W}$ and then $P^\tau = P_W^\tau$. The sequence $(E_t^{\mathrm{RIPR}})_{t\geq 1}$ is adapted to $\mathbf{F}$.

Sometimes (i.e., for some problem setting Π, $\mathbf{P}$, $\mathbf{Q}$ and choice of mixture or plug-in $\bar{Q}$) it is an e-process; sometimes not. If it is (as happens, e.g., in the tests described in Section 4.1.2), then $(E_t^{\mathrm{RIPR}})_{t\geq 1}$ is an e-process relative to an appropriately chosen $\bar{Q}$. It then dominates UI when using the same mixture over $\mathbf{Q}$: $E_t^{\mathrm{UI}} \leq E_t^{\mathrm{RIPR}}$, since they have the same numerator, but UI maximizes denominator likelihood. If it is not an e-process, then variations of RIPr can often still be used to obtain one. For example, in parametric k-sample tests (Section 4.2), $(E_t^{\mathrm{RIPR}})_{t\geq 1}$ is

an e-process when for $\bar{Q}$ we take any fixed $Q \in \mathbf{Q}$, but not when we take a plug-in or mixture $\bar{Q}$ that 'learns'. To handle composite $\mathbf{Q}$, we must then combine RIPr with the method of mixtures in a different way, as we now describe.

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, $E^R_\tau := \int E^\lambda_\tau \, dR(\lambda)$, 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. For example, in the two examples referred to above we have $\mathbf{Q} = \{Q_\theta : \theta \in \Theta_1\}$ a parametric set of distributions, and for each $\theta \in \Theta_1$, we define $M^\theta_t := (E^{\mathrm{RIPR}(Q_\theta)}_t)_{t \geq 1}$ to be the sequence of RIPr e-variables relative to point alternative Q_θ, which, as stated above, is an e-process. This is different from what we get if we try to use the method of mixtures directly to mix over $\mathbf{Q}$, defining some $\bar{Q}$ that 'learns' Q_θ: in that case $(E^{\mathrm{RIPR}(\bar{Q})}_t)_{t \geq 1}$ does not provide an e-process in the k-sample test of Section 4.2 below. Indeed, Section 4.1.2 successfully mixes over $\mathbf{Q}$, Sections 4.2 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 (a) every Bayes factor for a simple null also defines a test martingale. Further, (b) as pointed out by Darling and Robbins (1968), if X_i are i.i.d. from P, and P has a finite MGF, meaning $\Phi(\lambda) := \mathbb{E}_P[\exp(\lambda X_i)] < \infty$ for a given $\lambda > 0$, then $\exp(\lambda \sum_{i \leq t} X_i)\Phi(\lambda)^{-t}$ forms a test martingale for P. Construction (a) amounts to positing a specific alternative (a mixture of elements of $\mathbf{Q}$) and thus fits within both a Neyman–Pearson and Bayesian view on testing. In (b) there is no explicit alternative, and therefore it may perhaps be more in line with Fisher's view on testing.

Below we emphasize examples with composite nulls, for which by now a plethora of e-variables have been designed. Some of these examples (logrank test, t-test, contingency tables) are implemented in the R package `safestats` on CRAN (Turner et al., 2022).

4.1 Reduction to a Simple Null

One of the simplest practical ways to create e-processes with composite parametric nulls is to reduce the testing problem to one with a simple null, for example by coarsening or conditioning on a sufficient statistic of the data, so that the resulting marginal or conditional likelihood ratio has the same distribution under all $P \in \mathbf{P}$, making this likelihood ratio an e-process. This strategy has already been applied in the early days of sequential testing with Wald's sequential probability ratio test, which, while different from ours (Section 2.6), is nevertheless based on a likelihood ratio between a simple null and a simple alternative. This led to sequential versions of the t-test (by marginalization, in 1950) and other group invariant testing problems as well as to a sequential 2×2 contingency table test (by conditioning, in 1945). As reviewed in Section 4.1.2, it was recently discovered that for group invariant problems, such a reduction to a simple null leads to e-processes that have both GROW and REGROW status, and are thus highly suitable for our anytime-valid context. As reviewed in Section 4.1.3, the conditioning approach for the 2×2 table is suboptimal from a GRO perspective though. We review in Section 4.2 a general approach to e-processes for independence testing that does lead to an optimal e-process for the 2×2 table. But first we show that if H_0 and H_1 are separated, then optimal e-variables may sometimes simply be obtained by taking a likelihood ratio q/p_0 where p_0 represents a special element inside $\mathbf{P}$.

4.1.1 *Monotone LR families*; *logrank test*. Fix a 1-dimensional regular exponential family given in terms of its densities $\{p_\theta : \theta \in \Theta\}$ with Θ representing either its mean- or its natural parameterization. Fix $\theta_0 < \theta_1$ and suppose that under the null, the Y_i are i.i.d. $\sim P_\theta$ for $\theta \leq \theta_0$, whereas under the alternative, $\theta = \theta_1 > \theta_0$. An easy calculation (Grünwald, De Heide and Koolen, 2023, Example 4) shows that the GRO e-variable for this problem is given by $E^\theta_t := \prod_{i \leq t} p_\theta(Y_i)/p_{\theta_0}(Y_i)$, that is, it coincides with the likelihood ratio for a simple-vs.-simple testing problem. If the alternative is composite as well, that is, $\theta \geq \theta_1$ for some $\theta_1 > \theta_0$, we have a choice: the GROW e-variable ((15) with $\gamma = 0$) is given by setting $\theta = \theta_1$; alternatively, for a REGROW approach one can try to learn θ using the plug-in or mixture method as in Section 3.2.2. The crucial property needed for these results to hold is a *monotone likelihood ratio property*, which does not only hold for exponential families but also, for example, if p_θ represents a t-distribution with a fixed degree of freedom and noncentrality parameter θ.

It also extends beyond the i.i.d. case: Ter Schure et al. (2021) use this idea to provide efficiently computable GROW e-variables and test martingales for the logrank test, a work-horse of medical statistics. The test martingale they derive can be extended to the more general setting of the Cox regression model with covariates, for which computationally efficient implementation remains a work in progress. For the simple logrank test without covariates they provide extensive simulations showing that whether GRO is preferable over REGROW is a subtle matter. Here we focus on this simple logrank test. One

starts with m_0 subjects, partitioned into a *treatment a* and a *control* group b; for example, one wants to test a COVID vaccine; at time 0 all $m_{a,0}$ subjects in the treatment group get the vaccine and all $m_{b,0} = m - m_{a,0}$ subjects in the control get a placebo. At every time $t = 1, 2, \ldots$ an *event* (e.g., onset of covid) happens, either to somebody in the treatment ($Y_i = a$) or in the control ($Y_i = b$) group. Under the null, treating patients does not yield any benefit, so that $P(Y_i = a|Y^{i-1}) = \theta m_{a,i}/(\theta m_{a,i} + m_{b,i})$ for some $\theta \geq 1$ (if $\theta > 1$ then the treatment harms). Under the alternative, one assumes that this *hazard ratio* θ satisfies $\theta \leq \theta_{\max}$ for some $\theta_{\max} < 1$. Thus, at every i, one tests two separated sets of Bernoulli distributions, which can be addressed using the reduction to a null indicated above.

4.1.2 *t-test, regression, general group invariance.* Consider the following version of the t-test: according to the null, the X_t are i.i.d. $\sim N(\delta_0\sigma, \sigma)$ for some given effect size δ_0; according to the alternative, they are i.i.d. $N(\delta_1\sigma, \sigma)$ 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, $(V_t)_t$, by virtue of it being *scale-free* (it does not change if all data points are divided by a fixed constant), has a distribution P_{δ_0} that does not depend on the variance; similarly, under each distribution in the alternative, $(V_t)_t$ has the same marginal distribution P_{δ_1}. 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, $w(\sigma) = 1/\sigma$, that is,

$$E_t = \frac{\int_{\sigma>0} p_{\delta_1\sigma,\sigma}(X^t) w(\sigma)\, d\sigma}{\int_{\sigma>0} p_{\delta_0\sigma,\sigma}(X^t) w(\sigma)\, d\sigma}, \tag{18}$$

where $p_{\mu,\sigma}$ denotes the density of a $N(\mu, \sigma)$ distribution. Lai (1976a) also noted the equality (18) 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 Pérez-Ortiz et al. (2022) showed that, for all τ, E_τ is both the GROW and the REGROW e-variable: among all e-variables for the given null it maximizes

$$\inf_{\sigma>0} \mathbb{E}_{P_{\delta_1\sigma,\sigma}}\big(\log E_\tau - \gamma \log E_\tau^{\mathrm{RIPR}(P_{\delta_1\sigma,\sigma})}\big),$$

for every $\gamma \in \mathbb{R}$, including $\gamma = 0$ (GROW) and $\gamma = 1$ (REGROW), which is possible since

$$\mathbb{E}_{P_{\delta_1\sigma,\sigma}}\big(\log E_\tau^{\mathrm{RIPR}(P_{\delta_1\sigma,\sigma})}\big)$$

turns out to be constant in σ. Note that this is the appropriate generalization of the (RE)GROW criterion ((15) to the present case, with $\mathbf{Q}$ set to $\{P_{\delta_1,\sigma,\sigma} : \sigma > 0\}$ and since the null is now composite, M_τ^Q now set to $E_\tau^{\mathrm{RIPR}(P_{\delta_1\sigma,\sigma})}$, which is now the GRO e-variable relative to $Q = P_{\delta_1\sigma,\sigma}$ (Section 3.3). This implies that the proposed e-process is in fact optimal in a strong sense. They also demonstrate (RE)GROW optimality for the case (as in Section 4.1.1 above) that the defining constraint in either H_0 or H_1 or both is replaced by an inequality, that is, 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 (18) 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 composite settings we discuss in the next sub-section), test martingales overlap with a specific popular type of Bayes factors. Pérez-Ortiz et al. (2022) also extend these insights to the general setting of H_0 and H_1 that share nuisance parameters expressing a group invariance: under a mild regularity condition, the (RE)GROW e-variable turns out to be equivalent to a likelihood ratio for a coarsening of the data, under which the null becomes simple. This includes scale invariance such as in the t-test above (if one divides all data points by the same constant, the e-variable for the t-test above remains invariant), but also location, rotation, and other invariances. By considering the affine group, one can handle linear regression problems with Gaussian errors as well.

4.1.3 *Conditional e-values.* Suppose that, under the null, $Y \sim P_\theta, \theta \in \Theta$, and Z is a sufficient statistic for Y, so that the density of Y conditional on Z is given by the same $p(Y|Z)$, for all $\theta \in \Theta$. Then, for any conditional probability density $q(Y|Z)$, we have by the same reasoning as used already in Section 2.1 that $S := q(Y|Z)/p(Y|Z)$ is an e-variable conditional on Z, and hence also unconditionally, that is, for all $\theta \in \Theta$,

$$\mathbb{E}_{Y\sim P_\theta}[S|Z] = 1 \quad \text{and so} \quad \mathbb{E}_{(Z,Y)\sim P_\theta}[S] = 1. \tag{19}$$

Such e-variables were already used implicitly by Wald (1945) in his approach to sequential independence testing in the 2×2 contingency table, in which the null is the Bernoulli model. The idea can be extended to more general exponential family nulls, but as Hao et al. (2023) show these usually do not have (RE)GRO(W) status.

4.2 Parametric Alternative, Nonparametric Null

A second broad class of e-variables arises when under the null, outcome Y is independent of some observed covariate X (or more generally, the null says that some measure of dependence takes on a value at most δ); under the

alternative, there is dependence (or more generally, the dependence is stronger than δ' for some $\delta' > \rho$). For simplicity assume $Y, Y_1, Y_2, \ldots$ are i.i.d. under both null and alternative, and, for now, assume that $X, X_1, X_2, \ldots$ are also i.i.d. ("random design") according to some known distribution R, both under null and alternative. Then, with $\mathbf{P}$ and $\mathbf{Q}$ distributions for (X, Y), for all $P \in \mathbf{P}$, we have $P(X) = R(X)$, $P(Y|X) = P(Y)$, whereas under all $Q \in \mathbf{Q}$, $Q(X) = R(X)$ but $Q(Y|X) \neq Q(Y)$ with nonzero probability. Assuming that all $P \in \mathbf{P}$ have density p, then for any conditional density $q(y|x)$, we define

$$s_q(X, Y) := \frac{q(Y|X)}{\int q(Y|x)\, dR(x)}. \tag{20}$$

The random variable $S_q := s_q(X, Y)$ is trivially seen to be an e-variable for arbitrary q. Grünwald, Henzi and Lardy (2023) show that $\int q(Y|x)\, dR(x)$ is in fact the RIPr for the point alternative on (X, Y) under which $X \sim R$ and $Y|X \sim Q|R$, and as a consequence, S_q is the GRO e-variable for $\mathbf{P}$ as above and $\mathbf{Q} = \{Q\}$, the simple alternative corresponding to density q. For such simple $\mathbf{Q}$, it can be extended to a test martingale/e-process by setting $M_t := \prod_{i \le t} s_q(X_i, Y_i)$. For composite $\mathbf{Q}$, we can 'learn' the right q using the plug-in or mixture methods (Section 3.2.2). Related ideas can be found in Duan et al. (2020) and Shaer, Maman and Romano (2023).

This simple idea is surprisingly powerful: Grünwald, Henzi and Lardy (2023) extend it to design e-processes for *conditional* independence testing in which there is a third random variable Z and we test whether $P(Y|X, Z) = P(Y|Z)$ under the *model-X assumption* (Candès et al., 2018), which holds, for example, in randomized clinical trials. Turner, Ly and Grünwald (2021), Turner and Grünwald (2023a) provide another, quite different extension, without Z_i, in which X_i is not random. Rather, it may take on a finite number, say k, of values, and we observe k *data streams*. To illustrate, consider the case $k = 2$ with $X_i \in \{a, b\}$ binary, $X_i = a$ denoting that Y_i is an outcome in the 'treatment' group and $X_i = b$ that Y_i is in the 'control' group. We then observe streams $Y_{1,a}, Y_{2,a}, \ldots$ and $Y_{1,b}, Y_{2,b}, \ldots$. This is a sequential two-sample test (revisited in Section 5.9.3): the null now expresses that both streams are i.i.d. according to the same P, under the alternative, the distributions differ. Although now the X_i are not i.i.d. any more, we can 'flatten' the e-variables (20) so that they can still be used in this context. Turner, Ly and Grünwald (2021) show that variations of the resulting e-variables enjoy (RE)GROW optimality properties. Turner and Grünwald (2023a) further extend the setting to the case that the null does not express independence but rather that the strength of dependence is bounded by some user-supplied effect size δ (such as difference in mean or, if $Y_i \in \{0, 1\}$, the odds-ratio), and use this to develop anytime-valid confidence sequences. Finally, Turner and Grünwald (2023b) allow addition of *strata* Z thus providing a safe Cochran–Mantel–Haenszel test.

5. NONPARAMETRIC EXAMPLES

We present case studies that illustrate how one builds

(A) test martingales for composite nonparametric $\mathbf{P}$ despite there being no common reference measure,
(B) test *super*martingales when no martingales exist,
(C) e-processes when no supermartingales exist, and
(D) confidence sequences (CSs) for functionals using *reversed* submartingales instead of test martingales.

They are presented in an order that aids readability due to complexity of concepts, rather than (A) to (D) above. Many of the CSs have been implemented in C++ in the package `confseq`,[7] with Python and R interfaces.

5.1 Estimating Sub-Gaussian Means (Case B)

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

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

where $\mu_i := \mathbb{E}_P[X_i|\mathbf{F}_{i-1}]$. Fixing σ, let $\mathbf{G}^\mu$ be the set of sub-Gaussian distributions with parameter σ and $\mu_i = \mu$ for all i; this is a nonparametric generalization of Gaussianity. Darling and Robbins (1968) constructed CSs for μ. They observed that for any $\lambda \in \mathbb{R}$,

$$M_t^\mu(\lambda) := \exp\left(\lambda \sum_{i \le t}(X_i - \mu) - \frac{\lambda^2}{2}\sigma^2 t\right) \tag{21}$$

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

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

which is also a test supermartingale for $\mathbf{G}^\mu$. It grows exponentially fast under any $P \in \mathbf{G}^\theta$ for any $\theta \neq \mu$, and it grows faster for θ farther 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 (12), we find that

$$\frac{Y_t}{t} \pm \sigma\sqrt{\frac{(t\rho^2 + 1)}{t^2\rho^2}\log\big((t\rho^2 + 1)/\alpha^2\big)} \tag{22}$$

is a CS for μ. If μ_i differs for each i, an identical argument shows that (22) 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., 2021a, Proposition 2)). The $\sqrt{\log t}$ can be changed to $\sqrt{\log\log t}$ at

[7] https://github.com/gostevehoward/confseq.

the expense of other constants using mixture distributions that are unbounded at the origin; see Howard et al. (2021a, equation (1)).

Waudby-Smith and Ramdas (2023) observed that

$$\exp\Big(\sum_{i\le t}\Big(\lambda_i(X_i-\mu)-\frac{\lambda_i^2}{2}\sigma^2\Big)\Big)$$

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

5.2 Heavy-Tailed, Robust Mean Estimation (Case B)

For a fixed $\sigma > 0$, let $\mathbf{V}^\mu \supset \mathbf{G}^\mu$ be the set of distributions on $\mathbb{R}^\infty$ that yield observations with conditional mean μ and conditional variance bounded above by σ^2. Inspired by Catoni (2012), Wang and Ramdas (2023b) prove that

$$L_t^\mu := \exp\Big(\sum_{i\le t}\varphi\big(\lambda(X_i-\mu)\big)-\frac{\lambda^2}{2}\sigma^2 t\Big)$$

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 $x<0$. They then 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 can be relaxed with almost no practical consequence. The authors also derive extensions for the case when the pth moment is finite ($p>1$).

There are other known test supermartingales for this setting, for example by Dubins and Savage (1965) and Delyon (2009, Proposition 12) but experiments show that these to be significantly less powerful.

Wang and Ramdas (2023c) later extended these ideas to derive "Huber-robust" test supermartingales and CSs that can handle adversarial corruptions to heavy-tailed data.

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. 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. (2021a) prove that for any $\lambda\in[-1,1]$, and any predictable $\hat\mu_i\in\mathbf{F}_{i-1}$,

$$N_t^\mu(\lambda) := \exp\Big(\lambda\sum_{i\le t}(X_i-\mu)-\psi(\lambda)\sum_{i\le t}(X_i-\hat\mu_i)^2\Big), \tag{23}$$

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 (21), we can employ a fully empirical variance term in (23). 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 $\hat\mu_i$, obtained using some tricky algebra, is useful in lowering the empirical variance. Mixing over λ with a (conjugate) gamma distribution leads to a closed form mixture supermartingale; see Howard et al. (2021a) for the resulting CS.

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

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

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 $\bigcup_{\mu\in\mathbb{R}}\mathbf{B}^\mu$, 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

$$K_t^\mu := \prod_{i=1}^t\big(1+\lambda_i^\mu(X_i-\mu)\big), \tag{24}$$

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.

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

$t \geq 1$. Extending an older result by Efron (1969), de la Peña (1999, Lemma 6.1) establishes that for any $\lambda \in \mathbb{R}$,

$$R_t(\lambda) := \exp\Big(\lambda \sum_{i \leq t} X_i - \frac{\lambda^2}{2} \sum_{i \leq t} X_i^2\Big) \tag{25}$$

is a test supermartingale for $\mathbf{P}$. Notice again the fully empirical variance term, as in (23); 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 (5) are nonnegative and predictable, and $B_t - 1$ is an odd function of X_t. The unit bet underlying (25) takes the form $g(x) := \exp(\lambda x - \lambda^2 x^2/2)$. Since $g(x) - 1$ is not an odd function, mirroring it around one can improve it: using unit bets $\widetilde{g}(x) := g(x)1_{x \geq 0} + (2 - g(-x))1_{x \leq 0}$ yields a strictly better (and admissible) test martingale R_t^o.

5.5 Testing Exchangeability and Log-Concavity (Case C)

In the previous example, $\mathbf{P}$ is a 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. (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 even a (nontrivial) 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. Gangrade, Rinaldo and Ramdas (2023) prove that there is no test supermartingale for $\mathbf{L}_d$, but the universal inference approach yields a powerful e-process for $\mathbf{L}_d$.

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 $\phi : \Pi \mapsto \mathbb{R}_{\geq 0}$ is called convex if $\phi(aP + (1-a)Q) \leq a\phi(P) + (1-a)\phi(Q)$ for any $P, Q \in \Pi$ and $a \in [0,1]$. Classic examples are the entropy and the mean. Similarly, a divergence $D : \Pi \times \Pi \mapsto \mathbb{R}_{\geq 0}$ is called convex if $D(aP + (1-a)P', aQ + (1-a)Q') \leq aD(P,Q) + (1-a)D(P',Q')$. 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 $X_1, X_2, \ldots, X_t, \cdots \sim P$ and let P_t denote the empirical distribution of X^t. The exchangeable filtration $\mathcal{E}_t$ is the *decreasing* filtration given by $\mathcal{E}_t = \sigma(P_t, X_{t+1}, X_{t+2} \ldots)$; in words: $X_{t+1}, X_{t+2} \ldots$ are known perfectly, but the order of $X_1, X_2, \ldots, X_t$ 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; if this happens, call the time at which a change is proclaimed as 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 $\sup_{P\in\mathbf{P},\nu>0,Q\in\mathbf{Q}} \mathbb{E}_{P,\nu,Q}[\tau^* - \nu|\tau^* > \nu]$, 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

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

(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, $M_t := \sum_{i\le t} A_i$ is an e-detector for $\mathbf{P}$, where A_i is an e-process for $\mathbf{P}$ that depends only on $X_i, X_{i+1}, \ldots$. 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 $\tau^* := \inf\{t \ge 1 : M_\tau \ge 1/\alpha\}$, 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 (e.g.) the e-processes constructed earlier in this section. For many such nonparametric problems, e-detectors provide, as far as we know, the first change detection procedures with provable ARL control.

E-detectors only work when $\mathbf{P}$ and $\mathbf{Q}$ are prespecified and nonintersecting (e.g., we wish to detect mean changes from ≤ 0 to > 0). Shekhar and Ramdas (2023b) develop a complementary framework for change detection when only such information is not available (e.g., we wish to detect mean changes from anything to anything else) using a new notion of a "backward confidence sequence" (BCS). Here, one constructs two $(1-\alpha)$-confidence sequences—one forwards in time, and one backwards in time (i.e., forwards in time after reversing time)—and declares a changepoint the first time that they do not intersect. One can show that the ARL is controlled nonasymptotically, and in parametric settings one achieves an optimal detection delay up to constants. Due the plethora of CSs developed for nonparametric settings, plugging them into the BCS framework yields the first sequential change detection method for many nonparametric problems.

5.8 Time-Uniform Central Limit Theory and *Asymptotic* Confidence Sequences

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. For brevity, we focus here 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. For a sequential analog, a suitable generalization of the concept of confidence sequences is required, because (by definition) CSs have finite-sample validity guarantees that are already impossible in nonsequential settings.

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 $(C_t)_{t\ge1}$ is called an asymptotic CS if there exists some unknown nonasymptotic CS $(D_t)_{t\ge1}$, 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 (22) 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.

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 *Compendium of exponential supermartingales*. In previous sections, we have encountered several nonparametric test supermartingales of the form

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

where sum_t is the sum of the observed random variables, and variance_t captures their cumulative variance; recall (21), (23), (25) for example. Likelihood ratios for in exponential families also take an identical form, where sum_t adds the sufficient statistics. 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. 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 *Quantiles*. Howard and Ramdas (2022) derive confidence sequences based on i.i.d. 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) of any arbitrary univariate random variable, proving 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 other assumptions. This is one of the best studied problems in statistics. Nonparametric methods in the offline setting include the univariate Kolmogorov-Smirnov test and the multivariate kernel maximum mean discrepancy, amongst many others.

As to sequential two-sample tests, we should distinguish between three approaches. We already discussed the two-sample tests from Turner et al. (2023a, 2023b) in which the alternative is parametric in Section 4.2. Although we called these tests 'parametric', their null is better viewed as nonparametric. Second, in some works we can use arbitrary (e.g., deep learning based) sequential predictors that attempt to predict, given an outcome, from which of the two samples it was taken. These include Lhéritier and Cazals (2018) (although the likelihood ratio process they use technically is not an e-process) and Pandeva et al. (2022) (which may be re-interpreted as a modification of the above process based on UI, so that it does become an e-process);

Finally, one may directly attempt to obtain sequential analogs of large classes of existing offline nonparametric tests. Shekhar and Ramdas (2023a) successfully pursue this direction (even for non-i.i.d. data) in a general game-theoretic framework, and their e-processes perform excellently 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. Extending the above, the first sequential nonparametric independence testing framework was developed in Podkopaev et al. (2023), which allows random variables to lie in general spaces and also 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 $\{x_1, \ldots, x_N\}$, 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 (23)), and invert them to construct CSs. Waudby-Smith and Ramdas (2023) designed more powerful test martingales of the form (24) and the resulting confidence sequences are the tightest known so far. 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 4.1.1'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 ith 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 $S_{(1)}, S_{(2)}, \ldots$ 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

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

and $p_{(k)}$ represents the kth 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 nonnulls, 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

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

and $e_{[k]}$ represents the kth largest e-value. Surprisingly, this procedure controls the FDR at α under arbitrary dependence between the e-values. The same result holds if one picks any set of S e-values that are all larger than $K/(S\alpha)$; an analogous result does not hold for p-values.

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.

When both p-values and e-values are available for the same set of hypotheses (e.g., 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.

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 $1 \leq R \leq |S|$ 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).

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.

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 may be 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/Evaluating 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, Section 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, Section 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 commonly 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)).

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 4.1.2). 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 (2023)) 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 the density ratio $d\pi_0(\theta^*)/d\pi_t(\theta^*)$ is a test martingale for P_{θ^*}, termed the "prior-posterior ratio martingale" (it is used for different purposes in Bayesian statistics, where it is called the *Dickey-Savage ratio*. Thus, $\{\theta \in \Theta : d\pi_0(\theta)/d\pi_t(\theta) < 1/\alpha\}$ 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; we comment further in Appendix A. These are also closely related to Bayesian "snug regions" proposed[8] by Hildreth (1963).

8.1.2 *Group-sequential and alpha spending methods*. These methods 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: although there are exceptions[9] such as Mingxiu, Cappelleri and Gordon Lan (2007)—group sequential methods provide Type-I error control under multiple looks at the data, but they typically require a prespecified final sample size, and a prespecified set of times at which one looks at the data. In contrast, e-processes can be updated as long as new data is available; an extensive comparison in the setting of the logrank test is performed by Ter Schure, Grünwald and Ly (2021). 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 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 $q(x_t|x^{t-1})$ and with loss assessed by the *logarithmic loss function*

[8] We thank an anonymous reviewer for this reference.

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

$-\log q(x_t|x^{t-1})$. The vast field of *online learning* is about such sequential prediction and the logarithmic loss takes an important special place 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), Waudby-Smith and Ramdas (2023), Shekhar and Ramdas (2023a), Ramdas et al. (2022), Casgrain, Larsson and Ziegel (2022) and others. Importantly, whereas in this paper we emphasized the case that $\mathbf{Q}$ is a class of alternatives that are seriously contemplated as having generated the data, we may also do our tests with $\bar{q}$ that we suspect will predict the data reasonably well but cannot be considered 'potentially true' in any meaningful sense. For example, the $\bar{q}$ may just be an *expert*, a probabilistic predictor, the inner workings of which may be completely unknown to us.

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. A recent preprint by Zhang, Ramdas and Wang (2023) yields new insights via convex geometry and optimal transport.

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 coarsened filtration of conformal p-values. Despite the generality of conformal p-values, the coarsening implies safety under optional stopping for a smaller set of stopping times that cannot see the original data, but only the conformal p-values. This sacrifice is unnecessary for data with a known finite support, in which case an e-process is available in the original filtration, that appears in experiments to be at least as powerful (for binary data, or for data with small support) 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 (1976a) discussed in Section 4.1.2). 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 than (say) universal inference.

When can one coarsen 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) 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. Intriguingly, the use independent external randomization can significantly alter the story (Ramdas and Manole, 2023).

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, as opposed to just a sequence of e-variables? (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. (2023). In current applications the GRO-optimal RIPr e-variables are sometimes given by simple, analytic formulas (Sections 4.2 and 4.1.2), but for other applications numerical optimization is required (e.g., the logrank test as in Section 4.1.1). An algorithm by Li (1999) is slow. Do there exist practically effective algorithms?

We end by pointing the reader to Appendix B, where we discuss the question of whether there is a "price" to be paid by SAVI methods. We anticipate more discussion around this (part philosophical, part technical) topic.

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

E-processes can be seen as quantifying 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-processes 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, as exemplified by examples in the literature such as Pratt's volt-meter story (Edwards, 1992, Section 9.2) 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, admissible 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 4.1.2) 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 $\{E_\tau^\theta : \theta \in \Theta\}$ for data X^τ with

$$\bar{P}(\theta|X^\tau) := \frac{1}{E_\tau^\theta} = \frac{p_\theta(X^\tau)}{\int p_{\theta'}(X^\tau) w(\theta')\, d\theta'}$$

(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 $\{\theta : \bar{P}(\theta|\tau) > \alpha\}$, whenever the latter set is an interval. By Bayes' theorem, the Bayes posterior based on the same prior w is given by

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

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

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

We see that all elements θ of an e-confidence interval must have $\bar{P}(\theta|X^\tau) \geq \alpha$; 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.

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 $\{E_\tau^\theta : \theta \in \Theta\}$ 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 $\bar{P}(\theta|X^\tau) := 1/E_\tau^\theta$ 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 $\bar{P}(\theta|X^\tau)$ as a function of θ. Grünwald (2022, 2023) 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. 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 papers (see also Bates et al. (2022)), 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 no-one; 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: they allow to infuse prior knowledge into one's procedures; they output numbers with a clear evidential meaning; and they ensure error control and coverage.

APPENDIX B: DOES SAVI COME AT A PRICE?

We often receive questions like: "does the much greater flexibility allowed by SAVI methods compared to traditional Neyman–Pearson testing come at a price? Is there, for example, a loss of power or a need for larger samples before conclusions can be drawn? How competitive are SAVI methods with classical ones?"

Let us not be too starry-eyed here. There is always a fundamental information-theoretic tradeoff between different types of errors that is unavoidable by any framework, including ours.

At one level, one may view game-theoretic statistics as trading off the two errors differently from classical statistics—we ask for a more stringent form of "type-I error" that even holds under optional stopping or continuation of experiments, but accept a (slightly) higher "type-II error" in return, and many of the cited papers do detailed analysis on the quantitative nature of the aforementioned qualitative tradeoff, for example, Grünwald, De Heide and Koolen (2023), Ter Schure, Grünwald and Ly (2021), Waudby-Smith and Ramdas (2023), Howard et al. (2021a), Wang and Ramdas (2022).

However, the errors are in quotes because even our definitions of type-I and type-II errors themselves differ from classical statistics—instead of minimizing the probability of each error, when constructing e-statistics control corresponds to making sure the expectation of the e-statistic is at most one under the null, and low "type II" error corresponds to finding an e-statistic whose expected logarithmic value under the alternative is as large as possible.

Even talking about the "price" assumes that one set of metrics dominate the other, but we contend that our performance metrics are in fact more suitable in many situations, in that the resulting framework has a large number of benefits for reproducibility—even above the optional stopping/continuation, the methods are very robust to dependence (Section 6.2), it is easy to combine evidence from independent or dependent studies (Sections 2.9, 2.10 and 6.1), etc. Nevertheless, many SAVI papers directly and extensively compare 'real' Type-I and Type-II errors obtained with our approach (although that is not what we optimize for with the SAVI approach) to those obtained by classical frequentist approaches.

As to a comparison to Bayesian approaches, it is also difficult and of limited consequence to directly compare numbers, but in Appendix A we do give some feel about how Bayesian credible intervals compare to our anytime-valid confidence sets/intervals (which are wider than Bayesian credible intervals, but not wider than the support interval).

It is true that, if we compare an e-variable with GRO status for testing a specific H_0 and H_1 at a fixed sample size to the uniformly most powerful Neyman–Pearson test at that sample size (assuming it exists), the latter has more power. The exact difference (as well as, relatedly, difference of width between standard confidence intervals and anytime-valid confidence sequences, which tend to get between 1.5 and 2 times as wide) is investigated in detail, both theoretically and by simulation by the earlier cited papers (and many others).

On the other hand, if one allows for optional stopping and stops as soon as one can reject the null, then *expected* minimal stopping times under the alternative are about the same or even smaller than the fixed n needed to get the same power with a Neyman–Pearson test—illustrating that it all depends how one measures performance quality.

This suggests that (and indeed we feel that) really, the wrong questions are often being asked: the SAVI methods are optimizing for different criteria—a measure of evidence (wealth) that grows fast under any alternative, and does not grow under any null—which arguably are often more suitable in the applied sciences in which replicability issues abound. Thus, comparison in terms of power is only of limited use.

Grünwald (2022, 2023) also shows that based on e-processes, one has the flexibility to make decisions based on arbitrary loss functions that are determined post-hoc, which is arguably highly relevant for practice yet also is not captured by power.

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.

BARNARD, G. A. (1947). Review of Abraham Wald's *Sequential Analysis*. *J. Amer. Statist. Assoc.* **42** 658–665.

BARRON, A., RISSANEN, J. and YU, B. (1998). The Minimum Description Length principle in coding and modeling. *IEEE Trans. Inf. 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. Available at arXiv:2205.06812.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245 https://doi.org/10.1214/aos/1013699998

BENJAMINI, Y. and YEKUTIELI, D. (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *J. Amer. Statist. Assoc.* **100** 71–81.

BERGER, J. O., PERICCHI, L. R. and VARSHAVSKY, J. A. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhya, Ser. A* **60** 307–321. MR1718789

BREIMAN, L. (1961). Optimal gambling systems for favorable games. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 65–78. Univ. California Press, Berkeley, CA. MR0135630

CANDÈS, E., FAN, Y., JANSON, L. and LV, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 551–577. MR3798878 https://doi.org/10.1111/rssb.12265

CARNEY, D. R. My position on "Power Poses". Accessed 5 June 2022. Available at http://faculty.haas.berkeley.edu/dana_carney/pdf_my position on power poses.pdf.

CARNEY, D. R., CUDDY, A. J. C. and YAP, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychol. Sci.* **21** 1363–1368. https://doi.org/10.1177/0956797610383437

CASGRAIN, P., LARSSON, M. and ZIEGEL, J. (2022). Anytime-valid sequential testing for elicitable functionals via supermartingales. Available at arXiv:2204.05680.

CATONI, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185. MR3052407 https://doi.org/10.1214/11-AIHP454

CHOE, Y. J. and RAMDAS, A. (2023). Comparing sequential forecasters. *Oper. Res.* To appear. Available at arXiv:2110.00115.

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

COVER, T. M. (1974). Universal gambling schemes and the complexity measures of Kolmogorov and Chaitin. Technical Report, No. 12. Stanford Univ., Stanford, CA.

COX, D. R. (1952). Sequential tests for composite hypotheses. *Proc. Camb. Philos. Soc.* **48** 290–299. MR0047292 https://doi.org/10.1017/s030500410002764x

CRANE, H. and SHAFER, G. (2020). Risk is random: The magic of the d'Alembert. Available at: http://www.probabilityandfinance.com/articles/57.pdf.

DARLING, D. A. and ROBBINS, H. (1967). Confidence sequences for mean, variance, and median. *Proc. Natl. Acad. Sci. USA* **58** 66–68. MR0215406 https://doi.org/10.1073/pnas.58.1.66

DARLING, D. A. and ROBBINS, H. (1968). Some nonparametric sequential tests with power one. *Proc. Natl. Acad. Sci. USA* **61** 804–809. MR0238437 https://doi.org/10.1073/pnas.61.3.804

DAWID, A. P. (1984). Statistical theory. The prequential approach. *J. Roy. Statist. Soc. Ser. A* **147** 278–292. MR0763811 https://doi.org/10.2307/2981683

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. *Statist. Probab. Lett.* **81** 157–162. MR2740080 https://doi.org/10.1016/j.spl.2010.10.013

DE HEIDE, R. and GRÜNWALD, P. D. (2021). Why optional stopping can be a problem for Bayesians. *Psychon. Bull. Rev.* **28** 795–812. https://doi.org/10.3758/s13423-020-01803-x

DELYON, B. (2009). Exponential inequalities for sums of weakly dependent variables. *Electron. J. Probab.* **14** 752–779. MR2495559 https://doi.org/10.1214/EJP.v14-636

DE LA PEÑA, V. H. (1999). A general class of exponential inequalities for martingales and ratios. *Ann. Probab.* **27** 537–564. MR1681153 https://doi.org/10.1214/aop/1022677271

DIMITROV, V., SHAFER, G. and ZHANG, T. (2022). The martingale index. Available at: http://www.probabilityandfinance.com/articles/61.pdf.

DUAN, B., RAMDAS, A., BALAKRISHNAN, S. and WASSERMAN, L. (2020). Interactive martingale tests for the global null. *Electron. J. Stat.* **14** 4489–4551. MR4194269 https://doi.org/10.1214/20-EJS1790

DUAN, B., RAMDAS, A. and WASSERMAN, L. (2022). Interactive rank testing by betting. In *Proceedings First Conference on Causal Learning and Reasoning* PMLR.

DUBINS, L. E. and SAVAGE, L. J. (1965). A Tchebycheff-like inequality for stochastic processes. *Proc. Natl. Acad. Sci. USA* **53** 274–275. MR0182042 https://doi.org/10.1073/pnas.53.2.274

DUNN, R., RAMDAS, A., BALAKRISHNAN, S. and WASSERMAN, L. (2023). Gaussian universal likelihood ratio testing. *Biometrika* **110** 319–337. MR4588356 https://doi.org/10.1093/biomet/asac064

EDWARDS, A. W. F. (1992). *Likelihood*, Expanded ed. Johns Hopkins Univ. Press, Baltimore, MD. MR1191161

EFRON, B. (1969). Student's t-test under symmetry conditions. *J. Amer. Statist. Assoc.* **64** 1278–1302. MR0251826

FAN, X., GRAMA, I. and LIU, Q. (2015). Exponential inequalities for martingales with applications. *Electron. J. Probab.* **20** 1–22. MR3311214 https://doi.org/10.1214/EJP.v20-3496

FELLER, W. K. (1940). Statistical aspects of ESP. *J. Parapsychol.* **4** 271–298. MR0004461

GANGRADE, A., RINALDO, A. and RAMDAS, A. (2023). A sequential test for log-concavity. ArXiv preprint. Available at arXiv:2301.03542.

GRÜNWALD, P. (2022). Beyond Neyman–Pearson. Available at arXiv:2205.00901.

GRÜNWALD, P., DE HEIDE, R. and KOOLEN, W. (2023). Safe testing. *J. Roy. Statist. Soc. Ser. B.* To appear, with discussion.

GRÜNWALD, P., HENZI, A. and LARDY, T. (2023). Anytime-valid tests of conditional independence under model-X. *J. Amer. Statist. Assoc.*

GRÜNWALD, P. and ROOS, T. (2020). Minimum description length revisited. *Int. J. Math. Ind.* **11**.

GRÜNWALD, P. D. (2023). The e-posterior. *Philos. Trans. R. Soc. A* **381** 20220146. MR4590499

HAO, Y., GRÜNWALD, P., LARDY, T., LONG, L. and ADAMS, R. (2023). E-values for k-sample tests with exponential families. Available at arXiv:2303.0047.

HENDRIKS, H. (2018). Test martingales for bounded random variables. Available at arXiv:1801.09418.

HENZI, A., ARNOLD, S. and ZIEGEL, J. F. (2023). Sequentially valid tests for forecast calibration. *Ann. Appl. Stat.*

HENZI, A. and ZIEGEL, J. F. (2022). Valid sequential inference on probability forecast performance. *Biometrika* **109** 647–663.

HILDRETH, C. (1963). Bayesian statisticians and remote clients. *Econometrica* **31** 422–438.

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. MR4411508 https://doi.org/10.3150/21-bej1388

HOWARD, S. R., RAMDAS, A., MCAULIFFE, J. and SEKHON, J. (2020). Time-uniform Chernoff bounds via nonnegative supermartingales. *Probab. Surv.* **17** 257–317. MR4100718 https://doi.org/10.1214/18-PS321

HOWARD, S. R., RAMDAS, A., MCAULIFFE, J. and SEKHON, J. (2021a). Time-uniform, nonparametric, nonasymptotic confidence sequences. *Ann. Statist.* **49** 1055–1080. MR4255119 https://doi.org/10.1214/20-aos1991

IGNATIADIS, N., WANG, R. and RAMDAS, A. (2022). E-values as unnormalized weights in multiple testing. *Biometrika*. To appear. https://doi.org/10.1093/biomet/asad057

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. Clarendon Press, Oxford. MR0187257

JOHARI, R., KOOMEN, P., PEKELIS, L. and WALSH, D. (2022). Always valid inference: Continuous monitoring of A/B tests. *Oper. Res.* **70** 1806–1821. MR4451064 https://doi.org/10.1287/opre.2021.2135

JOHN, L. K., LOEWENSTEIN, G. and PRELEC, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* **23** 524–532. https://doi.org/10.1177/0956797611430953

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. *J. Mach. Learn. Res.* **22** 246. MR4353025

KELLY, J. L. JR. (1956). A new interpretation of information rate. *Bell Syst. Tech. J.* **35** 917–926. MR0090494 https://doi.org/10.1002/j.1538-7305.1956.tb03809.x

LAI, T. L. (1976a). On confidence sequences. *Ann. Statist.* **4** 265–280. MR0395103

LHÉRITIER, A. and CAZALS, F. (2018). A sequential non-parametric multivariate two-sample test. *IEEE Trans. Inf. Theory* **64** 3361–3370. MR3798382 https://doi.org/10.1109/TIT.2018.2800658

LI, J. Q. (1999). Estimation of Mixture Models Ph.D. thesis Yale Univ. New Haven, CT.

LI, J. Q. and BARRON, A. R. (2000). Mixture density estimation. In *Advances in Neural Information Processing Systems* **12** 279–285.

MACLEAN, L. C., THORP, E. O. and ZIEMBA, W. T. (2010). Long-term capital growth: The good and bad properties of the Kelly and fractional Kelly capital growth criteria. *Quant. Finance* **10** 681–687. MR2741943 https://doi.org/10.1080/14697688.2010.506108

MANOLE, T. and RAMDAS, A. (2023). Martingale methods for sequential estimation of convex functionals and divergences. *IEEE Trans. Inform. Theory*. **69** 4641–4658. MR4613565 https://doi.org/10.1109/TIT.2023.3250099

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. *Clin. Trials* **4** 329–340.

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. Available at arXiv:2110.14099.

PACE, L. and SALVAN, A. (2020). Likelihood, replicability and Robbins' confidence sequences. *Int. Stat. Rev.* **88** 599–615. MR4180669 https://doi.org/10.1111/insr.12355

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. Available at arXiv:2206.12290.

PÉREZ-ORTIZ, M. F., LARDY, T., DE HEIDE, R. and GRÜNWALD, P. (2022). E-statistics, group invariance and anytime valid testing. Available at 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. and MANOLE, T. (2023). Randomized and exchangeable improvements of Markov's, Chebyshev's and Chernoff's inequalities. ArXiv preprint. Available at arXiv:2304.02611.

RAMDAS, A., RUF, J., LARSSON, M. and KOOLEN, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. Available at arXiv:2009.03167.

RAMDAS, A., RUF, J., LARSSON, M. and KOOLEN, W. M. (2022). Testing exchangeability: Fork-convexity, supermartingales and e-processes. *Internat. J. Approx. Reason.* **141** 83–109. MR4364897 https://doi.org/10.1016/j.ijar.2021.06.017

RISSANEN, J. (1984). Universal coding, information, prediction, and estimation. *IEEE Trans. Inf. Theory* **30** 629–636. MR0755791 https://doi.org/10.1109/TIT.1984.1056936

ROBBINS, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* **58** 527–535. MR0050246 https://doi.org/10.1090/S0002-9904-1952-09620-8

ROBBINS, H. (1970). Statistical methods related to the law of the iterated logarithm. *Ann. Math. Stat.* **41** 1397–1409. MR0277063 https://doi.org/10.1214/aoms/1177696786

ROBBINS, H. and SIEGMUND, D. (1974). The expected sample size of some tests of power one. *Ann. Statist.* **2** 415–436. MR0448750

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. *Psychon. Bull. Rev.* **16** 225–237.

ROYALL, R. M. (1997). *Statistical Evidence*: *A Likelihood Paradigm*. *Monographs on Statistics and Applied Probability* **71**. CRC Press, London. MR1629481

RUF, J., LARSSON, M., KOOLEN, W. M. and RAMDAS, A. (2022). A composite generalization of Ville's martingale theorem. Available at arXiv:2203.04485.

RUSHTON, S. (1950). On a sequential t-test. *Biometrika* **37** 326–333. MR0044080 https://doi.org/10.1093/biomet/37.3-4.326

SHAER, S., MAMAN, G. and ROMANO, Y. (2023). Model-free sequential testing for conditional independence via testing by betting. In *International Conference on Artificial Intelligence and Statistics*.

SHAFER, G. (2021). Testing by betting: A strategy for statistical and scientific communication. *J. Roy. Statist. Soc. Ser. A* **184** 407–478. MR4255905 https://doi.org/10.1111/rssa.12647

SHAFER, G., SHEN, A., VERESHCHAGIN, N. and VOVK, V. (2011). Test martingales, Bayes factors and p-values. *Statist. Sci.* **26** 84–101. MR2849911 https://doi.org/10.1214/10-STS347

SHAFER, G. and VOVK, V. (2001a). *Probability and Finance*: *It's Only a Game! Wiley Series in Probability and Statistics*. *Financial Engineering Section*. Wiley Interscience, New York. MR1852450 https://doi.org/10.1002/0471249696

SHAFER, G. and VOVK, V. (2019). *Game-Theoretic Foundations for Probability and Finance*. Wiley, Hoboken, NJ.

SHEKHAR, S. and RAMDAS, A. (2023a). Nonparametric two sample testing by betting. *IEEE Trans. Inform. Theory*. To appear. https://doi.org/10.1109/TIT.2023.3305867

SHEKHAR, S. and RAMDAS, A. (2023b). Sequential change detection via backward confidence sequences. In *International Conference on Machine Learning*.

SHIN, J., RAMDAS, A. and RINALDO, A. (2022). E-detectors: A nonparametric framework for online changepoint detection. Available at 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. *F1000Res.* **11** 549. https://doi.org/10.12688/f1000research.74223.1

TER SCHURE, J., GRÜNWALD, P. and LY, A. (2021). Pandemic preparedness in data sharing; lessons learned from collaborating in a live meta-analysis. *STAtOR* **24** 47–52.

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. Available at arXiv:1906.07801.

TURING, A. M. (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. (2023a). Anytime-valid confidence intervals for contingency tables and beyond. *Statist. Probab. Lett* **198**.

TURNER, R. and GRÜNWALD, P. (2023b). Safe sequential testing and effect estimation in stratified count data. In *Annual AI and Statistics Conference*. PMLR.

TURNER, R., LY, A. and GRÜNWALD, P. (2021). Generic E-variables for exact sequential k-sample tests that allow for optional stopping. Available at 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, Paris.

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. *Statist. Sci.* **36** 595–611. MR4323055 https://doi.org/10.1214/20-sts817

VOVK, V., GAMMERMAN, A. and SHAFER, G. (2022). *Algorithmic Learning in a Random World*. Springer, Cham.

VOVK, V., NOURETDINOV, I. and GAMMERMAN, A. (2021). Conformal testing: Binary case with Markov alternatives. Available at arXiv:2111.01885.

VOVK, V. and WANG, R. (2021). E-values: Calibration, combination and applications. *Ann. Statist.* **49** 1736–1754. MR4298879 https://doi.org/10.1214/20-aos2020

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. *Ann. Math. Stat.* **16** 117–186. MR0013275 https://doi.org/10.1214/aoms/1177731118

WALD, A. (1947). *Sequential Analysis*. Wiley, New York. MR0020764

WANG, H. and RAMDAS, A. (2023a). The extended Ville's inequality for nonintegrable nonnegative supermartingales. ArXiv preprint. Available at arXiv:2304.01163.

WANG, H. and RAMDAS, A. (2023b). Catoni-style confidence sequences for heavy-tailed mean estimation. *Stochastic Process. Appl.* **163** 168–202. MR4610125 https://doi.org/10.1016/j.spa.2023.05.007

WANG, H. and RAMDAS, A. (2023c). Huber-robust confidence sequences. *26th International Conference on Artificial Intelligence and Statistics*.

WANG, R. and RAMDAS, A. (2022). False discovery rate control with e-values. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 822–852. MR4460577

WASSERMAN, L., RAMDAS, A. and BALAKRISHNAN, S. (2020). Universal inference. *Proc. Natl. Acad. Sci. USA* **117** 16880–16890. MR4242731 https://doi.org/10.1073/pnas.1922664117

WAUDBY-SMITH, I., ARBOUR, D., SINHA, R., KENNEDY, E. H. and RAMDAS, A. (2021). Time-uniform central limit theory and asymptotic confidence sequences. Available at arXiv:2103.06476.

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. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 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, Berlin.

WAUDBY-SMITH, I., WU, L., RAMDAS, A., KARAMPATZIAKIS, N. and MINEIRO, P. (2022). Anytime-valid off-policy inference for contextual bandits. ArXiv preprint. Available at 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. Available at arXiv:2203.12572.

ZHANG, Z., RAMDAS, A. and WANG, R. (2023). On the existence of powerful p-values and e-values for composite hypotheses. ArXiv preprint. Available at arXiv:2305.16539.