Estimating means of bounded random variables by betting

Ian Waudby-Smith[1] and Aaditya Ramdas[12]
Depts. of Stat. & Data Science[1], and Machine Learning[2], Carnegie Mellon Univ. †
E-mail: ianws@cmu.edu, aramdas@cmu.edu

[To be read before The Royal Statistical Society at a meeting organized by the Research Section on Tuesday, May 23rd, 2023, Professor Maria De Iorio in the Chair]

Abstract. This paper derives confidence intervals (CI) and time-uniform confidence sequences (CS) for the classical problem of estimating an unknown mean from bounded observations. We present a general approach for deriving concentration bounds, that can be seen as a generalization and improvement of the celebrated Chernoff method. At its heart, it is based on a class of composite nonnegative martingales, with strong connections to testing by betting and the method of mixtures. We show how to extend these ideas to sampling without replacement, another heavily studied problem. In all cases, our bounds are adaptive to the unknown variance, and empirically vastly outperform existing approaches based on Hoeffding or empirical Bernstein inequalities and their recent supermartingale generalizations by Howard et al. [2021]. In short, we establish a new state-of-the-art for four fundamental problems: CSs and CIs for bounded means, when sampling with and without replacement.

1. Introduction

This work presents a new approach to two fundamental problems: (Q1) how do we produce a confidence interval for the mean of a distribution with (known) bounded support using n independent observations? (Q2) given a fixed list of N (nonrandom) numbers with known bounds, how do we produce a confidence interval for their mean by sampling $n \leq N$ of them without replacement in a random order? We work in a nonasymptotic and nonparametric setting, meaning that we do not employ asymptotics or parametric assumptions. Both (Q1) and (Q2) are well studied questions in probability and statistics, but we bring new conceptual tools to bear, resulting in state-of-the-art solutions to both.

We also consider sequential versions of these problems where observations are made one-by-one; we derive time-uniform confidence sequences, or equivalently, confidence intervals that are valid at arbitrary stopping times. In fact, we first describe our techniques in the sequential regime, because the employed proof techniques naturally lend themselves to this setting. We then instantiate the derived bounds for the more familiar setting of a fixed sample size when a batch of data is observed all at once. Our supermartingale techniques can be thought of as generalizations of classical methods for deriving concentration inequalities, but we prefer to present them in the language of betting, since this is a more accurate reflection of the authors' intuition.

†Correspondence address: 5000 Forbes Ave, 132 Baker Hall, Pittsburgh, PA 15213, USA.

Arguably the most famous concentration inequality for bounded random variables was derived by Hoeffding [1963]. What is now referred to as "Hoeffding's inequality" was in fact improved upon in the same paper where he derived a Bernoulli-type upper bound on the moment generating function of bounded random variables [Hoeffding, 1963, Equation (3.4)]. While these bounds are already reasonably tight in a worst-case sense, the resulting confidence intervals do not adapt to non-Bernoulli distributions with lower variance. Inequalities by Bennett [1962], Bernstein [1927] and Bentkus [2004] improve upon Hoeffding's, but such improvements require knowledge of nontrivial upper bounds on the variance. This led to the development of so-called "empirical Bernstein inequalities" by Audibert et al. [2007]and Maurer and Pontil [2009], which outperform Hoeffding's method for low-variance distributions at large sample sizes by estimating the variance from the data. Our new, and arguably quite simple, approaches to developing bounds significantly outperform these past works (e.g. Figure 1).‡ We also show that the same conceptual (betting) framework extends to without-replacement sampling, resulting in significantly tighter bounds than classical ones by Serfling [1974], improvements by Bardenet and Maillard [2015] and previous state-of-the-art methods due to Waudby-Smith and Ramdas [2020].

For providing intuition, our approach can be described in words as follows: *If we are allowed to repeatedly bet against the mean being m, and if we make a lot of money in the process, then we can safely exclude m from the confidence set.* The rest of this paper makes the above claim more precise by showing smart, adaptive strategies for (automated) betting, quantifying the phrase "a lot of money", and explaining why such an exclusion is mathematically justified. At the risk of briefly losing the unacquainted reader, here is a slightly more detailed high-level description:

> For each $m \in [0, 1]$, we set up a "fair" multi-round game of statistician against nature whose payoff rules are such that if the true mean happened to equal m, then the statistician can neither gain nor lose wealth in expectation (their wealth in the m-th game is a nonnegative martingale), but if the mean is not m, then it is possible to bet smartly and make money. Each round involves the statistician making a bet on the next observation, nature revealing the observation and giving the appropriate (positive or negative) payoff to the statistician. The statistician then plays all these games (one for each m) in parallel, starting each with one unit of wealth, and possibly using a different, adaptive, betting strategy in each. The $1 - \alpha$ confidence set at time t consists of all $m \in [0, 1]$ such that the statistician's money in the corresponding game has not crossed $1/\alpha$. The true mean μ will be in this set with high probability.

Our choice of language above stems from a game-theoretic approach towards probability, as developed in the books by Shafer and Vovk [2001, 2019] and a recent paper by Shafer [2021], but from a purely mathematical viewpoint, our results are extensions of a unified supermartingale approach towards nonparametric concentra-

‡github.com/wannabesmith/betting-paper-simulations has code to reproduce figures. The `betting` module of the Python package in github.com/gostevehoward/confseq has the main algorithms, but the package also contains implementations from other papers.

tion and estimation described in Howard et al. [2020, 2021]; related supermartingale approaches were studied by Kaufmann and Koolen [2021], Jun and Orabona [2019]. We elaborate on this viewpoint in Section 4.1. The most directly related works to our own are by Hendriks [2018], whose preprint has initial explorations of methods similar to ours for with-replacement sequential testing and estimation, and Stark [2020], who credits Kaplan for a computationally intractable variant of our approach for sequential testing in the without-replacement case. Apart from several novel results, the present paper extends these past works in *depth, breadth and unity*: our work contains a deeper empirical and theoretical investigation from statistical and computational viewpoints, places our work in a broader context of related work in both settings, and unifies the with- and without-replacement methodology for both testing and estimation in both fixed-time and sequential settings.

We now have the appropriate context for a concrete formalization of our problem, which is slightly more general than introduced above. After that, we describe the game, why the rules of engagement result in valid statistical inference, and derive computationally and statistically efficient betting strategies.

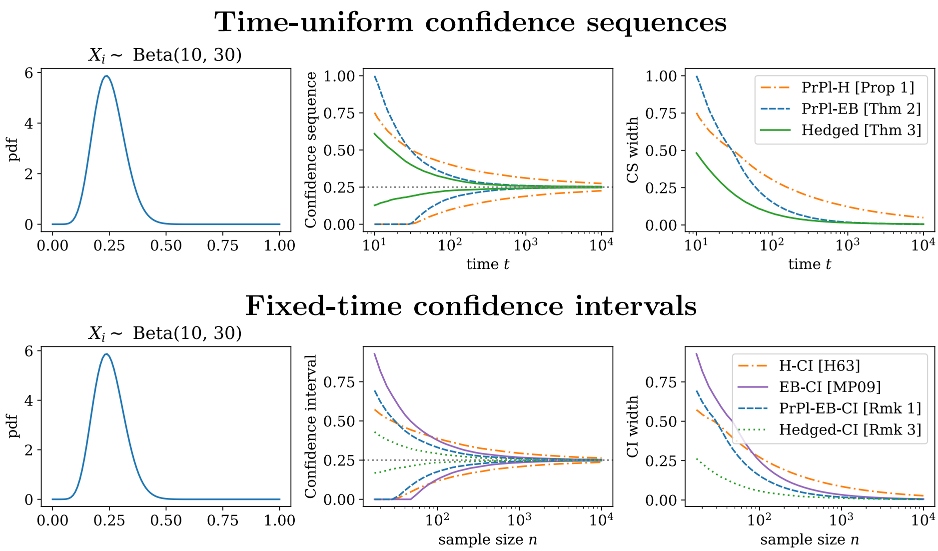

Figure 1. Time-uniform 95% confidence sequences (upper row) and fixed-time 95% confidence intervals (lower row) for the mean of independent and idenically distributed (iid) draws from a Beta(10, 30) distribution (unknown to the methods). The betting approaches (Hedged and Hedged-CI) adapt to both the small variance and asymmetry of the data, outperforming the other methods. For a detailed empirical comparison under a larger variety of settings, see Section C; for additional comparisons under non-iid data, see Section E.5.

Outline. We summarize the broad approach in Section 2. As a warmup, we derive a new predictable plug-in method for deriving confidence sequences using exponential supermartingales (Section 3), which already leads to computationally efficient and visually appealing empirical Bernstein confidence intervals and sequences. We then further improve on the aforementioned methods by developing a new martingale approach to deriving time-uniform and fixed-time confidence sets for means of bounded random variables, and connect the developed ideas to betting (Section 4). Section B discusses some principles to derive powerful betting strategies to obtain tight confidence sets. We then show how our techniques also extend to sampling without replacement (Section 5). Revealing simulations are performed along the way to demonstrate the efficacy of the new methods, with a more extensive comparison with past work in Section C. Section 6 summarizes how betting ideas have shaped mathematics, outside of our paper's focus on statistical inference. We postpone proofs to Section A and further theoretical insights to Section E.

2. Concentration inequalities via nonnegative supermartingales

To set the stage, let $\mathcal{Q}^m$ be the set of all distributions on $[0, 1]$, where each distribution has mean m. Note that $\mathcal{Q}^m$ is a convex set of distributions and it has no common dominating measure, since it consists of both discrete and continuous distributions.

Consider the setting where we observe a (potentially infinite) sequence of $[0, 1]$-valued random variables with conditional mean μ for some unknown $\mu \in [0, 1]$. We write this as $(X_t)_{t=1}^{\infty} \sim P$ for some $P \in \mathcal{P}^\mu$, where $\mathcal{P}^\mu$ is the set of all distributions P on $[0, 1]^\infty$ such that $\mathbb{E}_P(X_t \mid X_1, \ldots, X_{t-1}) = \mu$. This includes familiar settings such as independent observations, where $X_i \sim Q_i \in \mathcal{Q}^\mu$, or i.i.d. observations where all Q_i's are identical, but captures more general settings where the conditional distribution of X_t given the past is an element of $\mathcal{Q}^\mu$. When one only observes n outcomes, it suffices to imagine throwing away the rest, so that in what follows, we avoid new notation for distributions P over finite length sequences.

We are interested in deriving tight confidence sets for μ, typically intervals, with no further assumptions. Specifically, for a given error tolerance $\alpha \in (0, 1)$, a $(1 - \alpha)$ confidence interval (CI) is a random set $C_n \equiv C(X_1, \ldots, X_n) \subseteq [0, 1]$ such that

$$\forall n \geq 1, \inf_{P \in \mathcal{P}^\mu} P(\mu \in C_n) \geq 1 - \alpha. \tag{1}$$

As mentioned earlier, the inequality by Hoeffding [1963] implies that we can choose

$$C_n := \left(\overline{X}_n \pm \sqrt{\frac{\log(2/\alpha)}{2n}} \right) \cap [0, 1]. \tag{2}$$

Above, we write $(a \pm b)$ to mean $(a - b, a + b)$ for brevity.

This inequality is derived by what is now known as the Chernoff method [Boucheron et al., 2013], involving an analytic upper bound on the moment generating function of a bounded random variable. However, we will proceed differently; we adopt a hypothesis testing perspective, and couple it with a generalization of the Chernoff method. As mentioned in the introduction, we first consider the sequential regime where data are observed one after another over time, since nonnegative supermartingales — the primary mathematical tools used throughout this paper — naturally arise in this setup. As we will see, these sequential bounds can be instantiated for a fixed sample size, yielding tight confidence intervals for this more familiar setting.

These will be much tighter than the Hoeffding confidence interval (2), which is itself one such fixed-sample-size instantiation [Howard et al., 2020, Figures 4 and 6].

Let us briefly review some terminology. For succinctness, we use the notation $X_1^t := (X_1, \ldots, X_t)$. Define the sigma-field $\mathcal{F}_t := \sigma(X_1^t)$ generated by X_1^t with $\mathcal{F}_0$ being the trivial sigma-field. The *canonical filtration* $\mathcal{F} := (\mathcal{F}_t)_{t=0}^\infty$ refers to the increasing sequence of sigma-fields $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$. A stochastic process $(M_t)_{t=0}^\infty$ is called a *test supermartingale* for P if $(M_t)_{t=0}^\infty$ is a nonnegative process adapted to $\mathcal{F}$, $M_0 = 1$, and

$$\mathbb{E}_P(M_t \mid \mathcal{F}_{t-1}) \leq M_{t-1} \text{ for each } t \geq 1. \tag{3}$$

$(M_t)_{t=0}^\infty$ is called a *test martingale* for P if the above "$\leq$" is replaced with "$=$". We sometimes shorten $(M_t)_{t=0}^\infty$ to just (M_t) for brevity. If the above property holds simultaneously for all $P \in \mathcal{P}$, we call (M_t) a test (super)martingale for $\mathcal{P}$. We say that a sequence $(\lambda_t)_{t=1}^\infty$ is *predictable* if λ_t is $\mathcal{F}_{t-1}$-measurable for each $t \geq 1$, meaning λ_t can only depend on X_1^{t-1}. (In)equalities are interpreted in an almost sure sense.

2.1. Confidence sequences and the method(s) of mixtures

Even though the concentration inequalities thus far have been described in a setting where the sample size n is fixed in advance, all of our ideas stem from a sequential approach towards uncertainty quantification. The goal there is not to produce one confidence set C_n, but to produce an infinite sequence $(C_t)_{t=1}^\infty$ such that

$$\sup_{P \in \mathcal{P}^\mu} P(\exists t \geq 1 : \mu \notin C_t) \leq \alpha. \tag{4}$$

Such a $(C_t)_{t=1}^\infty$ is called a *confidence sequence* (CS), and preferably $\lim_{t\to\infty} C_t = \{\mu\}$. It is known [Howard et al., 2021, Lemma 3] that (4) is equivalent to requiring that $\sup_{P \in \mathcal{P}^\mu} P(\mu \notin C_\tau) \leq \alpha$ for arbitrary stopping times τ with respect to $\mathcal{F}$.

As detailed in the next subsection, one general way to construct a CS is to invert a family of sequential tests based on applying Ville's maximal inequality [Ville, 1939] to a test (super)martingale. In fact, Ramdas et al. [2020] proved that this is (in some formal sense) a universal method to construct CSs, meaning that any other approach can in principle be recovered or dominated by the aforementioned one.

Designing test supermartingales is nontrivial, and the task of making it have "power one" against composite alternatives is often accomplished via the *method of mixtures*. This can arguably be traced back (in a nonstochastic context) to Ville's 1939 thesis and (in a stochastic context) to Wald [1945]. Robbins and collabarators [Robbins and Siegmund, 1968, Robbins, 1970, Darling and Robbins, 1967a] applied the method to derive CSs, and these ideas have been extended to a variety of nonparametric settings by Howard et al. [2020, 2021]. The latter paper describes several variants: conjugate mixtures, discrete mixtures, stitching and inverted stitching.

These works form our vantage point for the rest of the paper, but we extend them in several ways. First, we describe a "predictable plug-in" technique that is implicit in the work of Ville. It can be viewed as a nonparametric extension of a passing remark in the parametric setting in the textbook by Wald [1945, Eq.10:10] and later explored in the parametric case by Robbins and Siegmund [1974].

Like Ville's work in the binary setting, the predictable plug-in method connects the game-theoretic approach and the aforementioned mixture methods — succinctly, the plugged-in value determines the bet, where each bet is implicitly targeting a different alternative (much like the components of a mixture). Following this translation, prior work on using the method mixtures for confidence sequences can be viewed as using the same betting strategy (mixture distribution) for every value of m. We find that there is significant statistical benefit to betting differently for each m (but tied together in a specific way, not in an ad hoc manner). One must typically specify the mixture distribution in advance of observing data, but betting can be viewed as building up a data-dependent mixture distribution on the fly (this led us to previously name our approach as the "predictable mixture" method). These sequential perspectives are powerful, even if only interested in fixed-sample CIs.

2.2. *Nonparametric confidence sequences via sequential testing*

As seen above, it is straightforward to derive a confidence interval for μ by resorting to a nonparametric concentration inequality like Hoeffding's. In contrast, it is also well known that CIs are inversions of families of hypothesis tests (as we will see below), so one could presumably derive CIs by first specifying tests. However, the literature on nonparametric concentration inequalities, such as Hoeffding's, has not commonly utilized a hypothesis testing perspective to derive concentration bounds; for example the excellent book on concentration by Boucheron, Lugosi, and Massart [2013] has no examples of such an approach. This is presumably because the underlying nonparametric, composite hypothesis tests may be quite challenging themselves, and one may not have nonasymptotically valid solutions or closed-form analytic expressions for these tests. This is in contrast to simple parametric nulls, where it is often easy to calculate a p-value based on likelihood ratios. In abandoning parametrics, and thus abandoning likelihood ratios, it may be unclear how to define a powerful test or calculate a nonasymptotically valid p-value. This is where betting and test (super)martingales come to the rescue. Ramdas et al. [2020, Proposition 4] prove that not only do likelihood ratios form test martingales, but every (nonparametric, composite) test martingale is also a (nonparametric, composite) likelihood ratio.

THEOREM 1 (4-STEP PROCEDURE FOR SUPERMARTINGALE CONFIDENCE SETS). *On observing $(X_t)_{t=1}^\infty \sim P$ from $P \in \mathcal{P}^\mu$ for some unknown $\mu \in [0,1]$, do*

(a) Consider the composite null hypothesis $H_0^m : P \in \mathcal{P}^m$ for each $m \in [0,1]$.

(b) For each index $m \in [0,1]$, construct a nonnegative process $M_t^m \equiv M^m(X_1, \ldots, X_t)$ such that the process $(M_t^\mu)_{t=0}^\infty$ indexed by μ has the following property: for each $P \in \mathcal{P}^\mu$, $(M_t^\mu)_{t=0}^\infty$ is upper-bounded by a test (super)martingale for P, possibly a different one for each P.

(c) For each $m \in [0,1]$ consider the sequential test $(\phi_t^m)_{t=1}^\infty$ defined by

$$\phi_t^m := \mathbf{1}(M_t^m \geq 1/\alpha),$$

where $\phi_t^m = 1$ represents a rejection of H_0^m after t observations.

(d) Define C_t as the set of $m \in [0,1]$ for which ϕ_t^m fails to reject H_0^m:

$$C_t := \{m \in [0,1] : \phi_t^m = 0\}.$$

Then $(C_t)_{t=1}^\infty$ is a $(1-\alpha)$-confidence sequence for μ: $\sup_{P \in \mathcal{P}^\mu} P(\exists t \geq 1 : \mu \notin C_t) \leq \alpha$.

The above result relies centrally on Ville's inequality [Ville, 1939], which states that if $(L_t) \equiv (L_t)_{t=1}^\infty$ is (upper bounded by) a test martingale for P, then we have $P(\exists t \geq 1 : L_t \geq 1/\alpha) \leq \alpha$. See [Howard et al., 2020, Section 6] for a short proof.

Proof (Theorem 1). By Ville's inequality, ϕ_t^m is a level-α sequential hypothesis test, in the sense that for any $P \in \mathcal{P}^\mu$, we have $P(\exists t \geq 1 : \phi_t^\mu = 1) \leq \alpha$. Now, by definition of the sets $(C_t)_{t=1}^\infty$, we have that $\mu \notin C_t$ at some time $t \geq 1$ if and only if there exists a time $t \geq 1$ such that $\phi_t^\mu = 1$, and hence

$$\sup_{P \in \mathcal{P}^\mu} P(\exists t \geq 1 : \mu \notin C_t) = \sup_{P \in \mathcal{P}^\mu} P(\exists t \geq 1 : \phi_t^\mu = 1) \leq \alpha, \tag{5}$$

which completes the proof. □

At a high level, this approach is not new. Composite test supermartingales for $\mathcal{P}$ have been used in past works on concentration inequalities and/or confidence sequences (which are related but different), from the initial series of works by Robbins and collaborators in the 1960s and 1970s, to de la Peña et al. [2007], to recent work by Jun and Orabona [2019, Section 7.2] and Howard et al. [2020, 2021]. Test martingales have also been explicitly considered in some hypothesis testing problems [Vovk et al., 2005, Shafer et al., 2011]; the latter paper popularized the term "test martingale" that we borrow, but unlike us, used it primarily for singleton $\mathcal{P} = \{P\}$. We highlight an (independently developed) unpublished preprint by Hendriks [2018] that has overlaps with the current paper in the with-replacement setting, and some complementary results. For singleton (parametric) classes $\mathcal{P}$, Wald's sequential likelihood ratio statistic is a test martingale, so all of the above methods can be viewed as inverting nonparametric or composite generalizations of Wald's tests.

Nevertheless, we make two additional comments. First, the requirement in step (b) of the algorithm that the process (M_t^m) be *upper-bounded* by a test (super)martingale for each $P \in \mathcal{P}$ was posited by Howard et al. [2020], and has recently been christened a e-process for $\mathcal{P}$ [Ramdas et al., 2021] (see also Grünwald et al. [2019]). E-processes are strictly more general than test (super)martingales for $\mathcal{P}$ in the sense that there exist many interesting classes $\mathcal{P}$ for which nontrivial test (super)martingales do not exist, but one can design powerful e-processes for $\mathcal{P}$. Second, one must take care to design test (super)martingales for each m that are tied together across m in a nontrivial manner that improves statistical power while maintaining computational tractability. All the confidence sets in this paper (both in the sequential and batch settings) will be based on this 4-step procedure, but with different carefully chosen processes (M_t^m). In the language of betting, we will come up with new, powerful ways to bet for each m, and also tie together the betting strategies for different m.

2.3. Connections to the Chernoff method

By virtue of $(C_t)_{t=1}^{\infty}$ being a time-uniform confidence sequence, we also have that C_n is a $(1-\alpha)$-confidence interval for μ for any fixed sample size n. In fact, the celebrated Chernoff method results in such a confidence interval. So, how exactly are the two approaches related? The answer is simple: Theorem 1 generalizes and improves on the Chernoff method. To elaborate, recall that Hoeffding proved that

$$\sup_{P \in \mathcal{P}^\mu} \mathbb{E}_P[\exp(\lambda(X-\mu) - \lambda^2/8)] \leq 1, \text{ for any } \lambda \in \mathbb{R}, \tag{6}$$

and so if X_1^n are independent (say), the following process can be used in Step (b):

$$M_t^m := \prod_{i=1}^{t} \exp\left(\lambda(X_i - m) - \lambda^2/8\right). \tag{7}$$

Usually, the only fact that matters for the Chernoff method is that $\mathbb{E}_P[M_t^m] \leq 1$, and Markov's inequality is applied (instead of Ville's) in Step (c). To complete the story, the Chernoff method then involves a smart choice for λ. Setting $\lambda := \sqrt{8\log(1/\alpha)/n}$ recovers the familiar Hoeffding inequality for the batch sample-size setting. Taking a union bound over X_1^n and $-X_1^n$ yields the Hoeffding confidence interval (2) exactly. Using our 4-step approach, the resulting confidence sequence is a time-uniform generalization of Hoeffding's inequality, recovering the latter precisely including constants at time n; see Howard et al. [2020] for this and other generalizations.

In recent parlance, a statistic like M_t^m, which has at most unit expectation under the null, has been called a betting score [Shafer, 2021] or an e-value [Vovk, 2021] and their relationship to sequential testing [Grünwald et al., 2019] and estimation [Ramdas et al., 2020] as an alternative to p-values has been recently examined. In parametric settings with singleton nulls and alternative hypotheses, the likelihood ratio is an e-value. For composite null testing, the split likelihood ratio statistic [Wasserman et al., 2020] (and its variants) are e-values. However, our setup is more complex: $\mathcal{P}^m$ is highly composite, there is no common dominating measure to define likelihood ratios, but Hoeffding's result yields an e-value. (In fact, it yields test supermartingale and hence an e-process, which is an e-value even at stopping times.)

In summary, the Chernoff method is simply one powerful, but as it turns out, rather limited way to construct an e-value. This paper provides better constructions of M_t^m, whose expectation is exactly equal to one, thus removing one source of looseness in the Hoeffding-type approach above, as well as better ways to pick the tuning parameter λ, which will correspond to our bet.

3. Warmup: exponential supermartingales and predictable plug-ins

A central technique for constructing confidence sequences (CSs) is Robbins' *method of mixtures* [Robbins, 1970], see also Darling and Robbins [1967a], Robbins and Siegmund [1968, 1970, 1972, 1974]. Related ideas of "pseudo-maximization" or Laplace's method were further popularized and extended by de la Peña et al. [2004, 2007, 2009], and has led to several other followup works [Abbasi-Yadkori et al., 2011, Balsubramani, 2014, Howard et al., 2020, Kaufmann and Koolen, 2021].

However, beyond the case when the data are (sub)-Gaussian, the method of mixtures rarely leads to a closed-form CS; it yields an *implicit* construction for C_t which can sometimes be computed efficiently (e.g. using conjugate mixtures [Howard et al., 2021]), but is otherwise analytically opaque and computationally tedious. Below, we provide an alternative construction — called the "predictable plug-in" — that is exact, explicit and efficient (computationally and statistically).

In the next section, our CSs avoid exponential supermartingales, and are much tighter than the recent state-of-the-art in Howard et al. [2021]. The ones in this section match the latter but are simpler to compute, so we present them first.

3.1. Predictable plug-in Cramer-Chernoff supermartingales

Suppose $(X_t)_{t=1}^{\infty} \sim P$ for some $P \in \mathcal{P}^{\mu}$ where $\mathcal{P}^{\mu}$ is the set of all distributions on $\prod_{i=1}^{\infty}[0,1]$ so that $\mathbb{E}_P(X_t \mid \mathcal{F}_{t-1}) = \mu$ for each t. The Hoeffding process $(M_t^H(m))_{t=0}^{\infty}$ for a given candidate mean $m \in [0,1]$ is given by

$$M_t^H(m) := \prod_{i=1}^{t} \exp\left(\lambda(X_i - m) - \psi_H(\lambda)\right) \tag{8}$$

with $M_0^H(m) \equiv 1$ by convention. Here $\psi_H(\lambda) := \lambda^2/8$ is an upper bound on the cumulant generating function (CGF) for $[0,1]$-valued random variables with $\lambda \in \mathbb{R}$ chosen in some strategic way. For example, to maximize $M_n^H(m)$ at a fixed sample size n, one would set $\lambda := \sqrt{8\log(1/\alpha)/n}$ as in the classical fixed-time Hoeffding inequality [Hoeffding, 1963].

Following Howard et al. [2021], we have that $(M_t^H(\mu))_{t=0}^{\infty}$ is a nonnegative supermartingale with respect to the canonical filtration. Therefore, by Ville's maximal inequality for nonnegative supermartingales [Ville, 1939, Howard et al., 2020],

$$P\left(\exists t \geq 1 : M_t^H(\mu) \geq 1/\alpha\right) \leq \alpha. \tag{9}$$

Robbins' method of mixtures proceeds by noting that $\int_{\lambda \in \mathbb{R}} M_t^H(m) dF(\lambda)$ is also a supermartingale for any "mixing" probability distribution $F(\lambda)$ on $\mathbb{R}$ and thus

$$P\left(\exists t \geq 1 : \int_{\lambda \in \mathbb{R}} M_t^H(\mu) dF(\lambda) \geq 1/\alpha\right) \leq \alpha. \tag{10}$$

In this particular case, if $F(\lambda)$ is taken to be the Gaussian distribution, then the above integral can be computed in closed-form [Howard et al., 2020]. For other distributions or altogether different supermartingales (i.e. other than Hoeffding), the integral may be computationally tedious or intractable.

To combat this, instead of fixing $\lambda \in \mathbb{R}$ or integrating over it, consider constructing a sequence $\lambda_1, \lambda_2, \ldots$ which is predictable, and thus λ_t can depend on X_1^{t-1}. Then,

$$M_t^{\text{PrPl-H}}(m) := \prod_{i=1}^{t} \exp(\lambda_i(X_i - m) - \psi_H(\lambda_i)) \tag{11}$$

is also a test supermartingale for $\mathcal{P}^m$ (and hence Ville's inequality applies). We call such a sequence $(\lambda_t)_{t=1}^{\infty}$ a *predictable plug-in*. While not always explicitly referred to by this exact name, predictable plug-ins have appeared in works on parametric sequential analysis by Wald [1947, Eq. (10:10)], Robbins and Siegmund [1974, Eq. (4)], Dawid [1984], and Lorden and Pollak [2005] as well as in the information theory literature [Rissanen, 1984]. As we will see, these techniques also prove useful in nonparametric testing and estimation problems both in sequential and batch settings.

Using $M_t^{\text{PrPl-H}}(m)$ as the process in Step (b) of Theorem 1 results in a lower CS for μ, while constructing an analogous supermartingale using $(-X_t)_{t=1}^\infty$ yields an upper CS. Combining these by taking a union bound results in the predictable plug-in Hoeffding CS which we introduce now.

PROPOSITION 1 (PREDICTABLE PLUG-IN HOEFFDING CS **[PrPl-H]**). *Suppose that* $(X_t)_{t=1}^\infty \sim P$ *for some* $P \in \mathcal{P}^\mu$. *For any chosen real-valued predictable* $(\lambda_t)_{t=1}^\infty$,

$$C_t^{\text{PrPl-H}} := \left(\frac{\sum_{i=1}^t \lambda_i X_i}{\sum_{i=1}^t \lambda_i} \pm \frac{\log(2/\alpha) + \sum_{i=1}^t \psi_H(\lambda_i)}{\sum_{i=1}^t \lambda_i} \right) \quad \textit{forms a } (1-\alpha)\textit{-CS for } \mu,$$

as does its running intersection, $\bigcap_{i \le t} C_i^H$.

A sensible choice of predictable plug-in is given by

$$\lambda_t^{\text{PrPl-H}} := \sqrt{\frac{8\log(2/\alpha)}{t\log(t+1)}} \wedge 1, \tag{12}$$

for reasons which will be discussed in Section 3.3. The proof of Proposition 1 is provided in Section A.1. As alluded to earlier, predictable plug-ins are actually the *least* interesting when using Hoeffding's sub-Gaussian bound because of the available closed form Gaussian-mixture boundary. However, the story becomes more interesting when either (a) the method of mixtures is computationally opaque or complex, or (b) the optimal choice of λ is based on unknown but estimable quantities. Both (a) and (b) are issues that arise when computing empirical Bernstein-type CSs and CIs. In the following section, we present predictable plug-in empirical Bernstein-type CSs and CIs which are both computationally and statistically efficient.

3.2. Application: closed-form empirical Bernstein confidence sets

To prepare for the results that follow, consider the empirical Bernstein-type process,

$$M_t^{\text{PrPl-EB}}(m) := \prod_{i=1}^t \exp\left\{\lambda_i(X_i - m) - v_i\psi_E(\lambda_i)\right\} \tag{13}$$

where, following Howard et al. [2020, 2021], we have defined $v_i := 4(X_i - \widehat{\mu}_{i-1})^2$ and

$$\psi_E(\lambda) := (-\log(1-\lambda) - \lambda)/4 \quad \text{for } \lambda \in [0,1). \tag{14}$$

As we revisit later, the appearance of the constant 4 is to facilitate easy comparison to ψ_H, since $\lim_{\lambda \to 0^+} \psi_E(\lambda)/\psi_H(\lambda) = 1$. In short, ψ_E is nonnegative, increasing on $[0,1)$, and grows quadratically near 0.

Using $M_t^{\text{PrPl-EB}}(m)$ in Step (b) in Theorem 1 — and applying the same procedure but with $(X_t)_{t=1}^\infty$ and m replaced by $(-X_t)_{t=1}^\infty$ and $-m$ combined with a union bound over the resulting CSs — we get the following CS.

THEOREM 2 (PREDICTABLE PLUG-IN EMPIRICAL BERNSTEIN CS **[PrPl-EB]**). *Suppose $(X_t)_{t=1}^\infty \sim P$ for some $P \in \mathcal{P}^\mu$. For any $(0,1)$-valued predictable $(\lambda_t)_{t=1}^\infty$,*

$$C_t^{\text{PrPl-EB}} := \left(\frac{\sum_{i=1}^t \lambda_i X_i}{\sum_{i=1}^t \lambda_i} \pm \frac{\log(2/\alpha) + \sum_{i=1}^t v_i \psi_E(\lambda_i)}{\sum_{i=1}^t \lambda_i} \right) \quad \textit{forms a } (1-\alpha)\textit{-CS for } \mu,$$

as does its running intersection, $\bigcap_{i \le t} C_i^{\text{PrPl-EB}}$.

In particular, we recommend the predictable plug-in $(\lambda_t^{\text{PrPl-EB}})_{t=1}^\infty$ given by

$$\lambda_t^{\text{PrPl-EB}} := \sqrt{\frac{2\log(2/\alpha)}{\widehat{\sigma}_{t-1}^2 t \log(1+t)}} \wedge c, \quad \widehat{\sigma}_t^2 := \frac{\frac{1}{4} + \sum_{i=1}^t (X_i - \widehat{\mu}_i)^2}{t+1}, \quad \widehat{\mu}_t := \frac{\frac{1}{2} + \sum_{i=1}^t X_i}{t+1} \quad (15)$$

for some $c \in (0,1)$ (a reasonable default being $1/2$ or $3/4$). This choice was inspired by the fixed-time empirical Bernstein as well as the widths of time-uniform CSs (more details are provided in Section 3.3). The sequences of estimators $(\widehat{\mu}_t)_{t=1}^\infty$ and $(\widehat{\sigma}_t^2)_{t=1}^\infty$ can be interpreted as predictable, regularized sample means and variances. This technique was employed by Kotłowski et al. [2010] for misspecified exponential families in the so-called *maximum likelihood plug-in strategy*.

The proof of Theorem 2 relies on establishing that $M_t^{\text{PrPl-EB}}(m)$ is a test supermartingale for $\mathcal{P}^m$. This latter fact is related to, but cannot be derived directly from, a powerful deterministic inequality for bounded numbers due to Fan et al. [2015]. One needs an additional trick from Howard et al. [2021, Section A.8] which swaps $(X_i - m)^2$ with $(X_i - \widehat{\mu}_{i-1})^2$, for any predictable $\widehat{\mu}_{i-1}$, within the variance term v_i. It is this additional piece which yields both tighter and *closed-form* CSs; details are in Section A.2. We remark that before taking the running intersection, the above intervals are symmetric around the weighted sample mean, but this symmetry will not carry forward to other CSs in the paper.

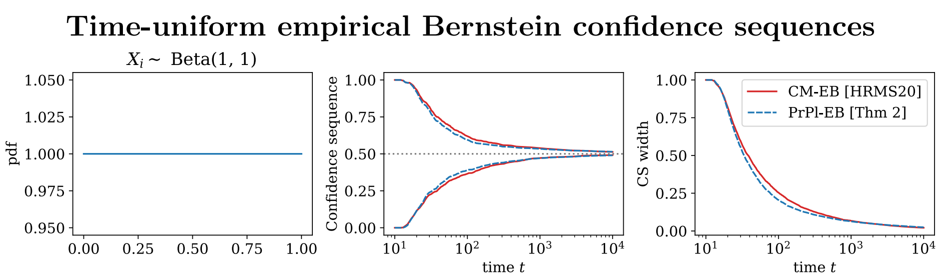

Figure 2. Empirical Bernstein CSs produced via a predictable plug-in (PrPl) with $(\lambda_t)_{t=1}^\infty$ from (15) match (or slightly improve) those obtained via conjugate mixtures (CM) by Howard et al. [2021]; the former is closed-form, but the latter is not and requires numerical methods.

Figure 2 compares the conjugate mixture empirical-Bernstein CS (CM-EB) due to Howard et al. [2021] with our predictable plug-in empirical-Bernstein CS (PrPl-EB). The two CSs perform similarly, but our closed-form PrPl-EB is over 500 times faster to compute than CM-EB (in our experience) which requires root finding at each step. However, our later bounds will be tighter than both of these.

REMARK 1. *Theorem 2 yields computationally and statistically efficient empirical Bernstein-type CIs for a fixed sample size n. Recalling (15), we recommend using $\bigcap_{i \le n} C_i^{\text{PrPl-EB}}$ along with the predictable sequence*

$$\lambda_t^{\text{PrPl-EB}(n)} := \sqrt{\frac{2\log(2/\alpha)}{n\widehat{\sigma}_{t-1}^2}} \wedge c. \tag{16}$$

We call the resulting confidence interval the "predictable plug-in empirical Bernstein confidence interval" or ***[PrPl-EB-CI]*** *for short; see Figure 3.*

If $X_1, \dots, X_n$ are independent, then at the expense of computation, the above CI can be effectively derandomized to remove the effect of the ordering of variables. One can randomly permute the data B times to obtain $(\widetilde{X}_{1,b}, \dots, \widetilde{X}_{n,b})$ and correspondingly compute $\widetilde{M}_{n,b}^{\text{PrPl-EB}}(m)$, one for each permutation $b \in \{1, \dots, B\}$. Averaging over these permutations, define $\widetilde{M}_n^{\text{PrPl-EB}}(m) := \frac{1}{B}\sum_{b=1}^B \widetilde{M}_{n,b}^{\text{PrPl-EB}}(m)$. For each b, $M_{n,b}^{\text{PrPl-EB}}(\mu)$ has expectation at most one (by linearity of expectation). Thus, $\widetilde{M}_n^{\text{PrPl-EB}}(\mu)$ is a e-value (i.e. it has expectation at most 1). By Markov's inequality, $\widetilde{C}_n^{\text{PrPl-EB}} := \{m \in [0,1] : \widetilde{M}_n^{\text{PrPl-EB}}(m) < 1/\alpha\}$ is a $(1-\alpha)$-CI for μ. This set is not available in closed-form and the intersection $\bigcap_{i \le n} \widetilde{C}_i^{\text{PrPl-EB}}$ no longer yield a valid CI. In our experience, this derandomization procedure neither helps nor hurts. In any case, both $\bigcap_{i \le n} C_i$ and $\widetilde{C}_n$ will be significantly improved in Section 4.4.

In Section E.3, we show that in iid settings the width of [PrPl-EB-CI] scales with the true (unknown) standard deviation:

$$\sqrt{n}\left(\frac{\log(2/\alpha) + \sum_{i=1}^n v_i \psi_E(\lambda_i)}{\sum_{i=1}^n \lambda_i}\right) \xrightarrow{a.s.} \sigma\sqrt{2\log(2/\alpha)}. \tag{17}$$

Notice that (17) is the same asymptotic behavior that one would observe for CIs based on Bernstein's or Bennett's inequalities, both of which require knowledge of the true variance σ^2, while [PrPl-EB-CI] does not. This is in contrast to the empirical Bernstein CIs of Maurer and Pontil [2009] whose limit would be $\sigma\sqrt{2\log(4/\alpha)}$. In the maximum variance case where $\sigma = 1/2$, (17) yields the same asymptotic behavior as Hoeffding's CI (2).

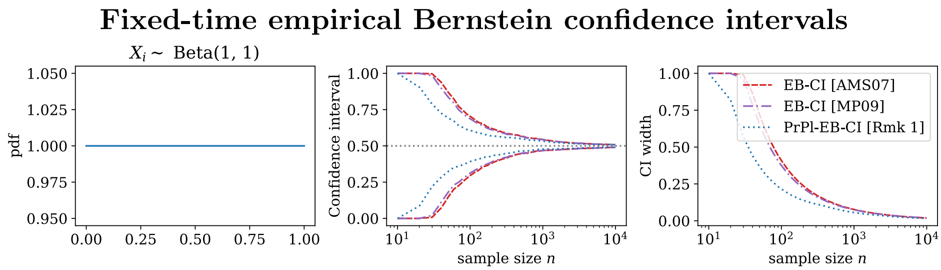

Figure 3. Our predictable plug-in (PrPl) empirical Bernstein (EB) CI is significantly tighter than those of Maurer and Pontil [2009] and Audibert et al. [2007].

Table 1. Below, we think of $\log x$ as $\log(x+1)$ to avoid trivialities. The claimed rates are easily checked by approximating the sums as integrals, and taking derivatives. For example, $\frac{d}{dx}\log\log x = 1/x\log x$, so the sum of $\sum_{i\le t} 1/i\log i \asymp \log\log t$. It is worth remarking that for $t = 10^{80}$, the number of atoms in the universe, $\log\log t \approx 5.2$, which is why we treat $\log\log t$ as a constant when expressing the rate for W_t. The iterated logarithm pattern in the the last two lines can be continued indefinitely.

Strategy $(\lambda_i)_{i=1}^\infty$	$\sum_{i=1}^t \lambda_i$	$\sum_{i=1}^t \lambda_i^2$	Width W_t
$\asymp 1/i$	$\asymp \log t$	$\asymp 1$	$1/\log t$
$\asymp \sqrt{\log i/i}$	$\asymp \sqrt{t}\log t$	$\asymp \log^2 t$	$\asymp \log^{3/2} t/\sqrt{t}$
$\asymp 1/\sqrt{i}$	$\asymp \sqrt{t}$	$\asymp \log t$	$\asymp \log t/\sqrt{t}$
$\asymp 1/\sqrt{i\log i}$	$\asymp \sqrt{t/\log t}$	$\asymp \log\log t$	$\asymp \sqrt{\log t/t}$
$\asymp 1/\sqrt{i\log i\log\log i}$	$\asymp \sqrt{t/\log t}$	$\asymp \log\log\log t$	$\asymp \sqrt{\log t/t}$

Until now, we presented various predictable plug-ins — $(\lambda_t^{\text{PrPl-H}})_{t=1}^\infty$, $(\lambda_t^{\text{PrPl-EB}})_{t=1}^\infty$, and $(\lambda_t^{\text{PrPl-EB}(n)})_{t=1}^n$ — but have not provided intuition for why these are sensible choices. Next, we discuss guiding principles for deriving predictable plug-ins.

3.3. *Guiding principles for deriving predictable plug-ins*

Let us begin our discussion with the predictable plug-in Hoeffding process (11) and the resulting CS in Proposition 1, which has a half-width

$$W_t = \frac{\log(2/\alpha) + \sum_{i=1}^t \lambda_i^2/8}{\sum_{i=1}^t \lambda_i}$$

To ensure that $W_t \to 0$ as $t \to \infty$, it is clear that we want $\lambda_t \xrightarrow{\text{a.s.}} 0$, but at what rate? As a sensible default, we recommend setting $\lambda_t \asymp 1/\sqrt{t\log t}$ so that $W_t = \widetilde{O}(\sqrt{\log t/t})$ which matches the width of the conjugate mixture Hoeffding CS [Howard et al., 2020, Proposition 2] (here $\widetilde{O}$ treats $O(\log\log t)$ factors as constants). See Table 3.3 for a comparison between rates for λ_t and their resulting CS widths.

Now consider the predictable plug-in empirical Bernstein process (13) and the resulting CS of Theorem 2, which has a half-width

$$W_t = \frac{\log(2/\alpha) + \sum_{i=1}^t 4(X_i - \widehat{\mu}_{i-1})^2\psi_E(\lambda_i)}{\sum_{i=1}^t \lambda_i}$$

By two applications of L'Hôpital's rule, we have that

$$\frac{\psi_E(\lambda)}{\psi_H(\lambda)} \xrightarrow{\lambda\to 0^+} 1. \tag{18}$$

Performing some approximations for small λ_i to help guide our choice of $(\lambda_t)_{t=1}^\infty$ (without compromising validity of resulting confidence sets) we have that

$$W_t \approx \frac{\log(2/\alpha) + \sum_{i=1}^t 4(X_i - \mu)^2\lambda_i^2/8}{\sum_{i=1}^t \lambda_i}. \tag{19}$$

Thus, in the special case of i.i.d. X_i with variance σ^2, for large enough t,

$$\mathbb{E}_P(W_t \mid \mathcal{F}_{t-1}) \lesssim \frac{\log(2/\alpha) + \sigma^2 \sum_{i=1}^t \lambda_i^2/2}{\sum_{i=1}^t \lambda_i}. \tag{20}$$

If we were to set $\lambda_1 = \lambda_2 = \cdots = \lambda^\star \in \mathbb{R}$ and minimize the above expression for a specific time $t^\star$, this amounts to minimizing

$$\frac{\log(2/\alpha) + \sigma^2 t^\star {\lambda^\star}^2/2}{t^\star \lambda^\star}, \tag{21}$$

which is achieved by setting

$$\lambda^\star := \sqrt{\frac{2\log(2/\alpha)}{\sigma^2 t^\star}}. \tag{22}$$

This is precisely why we suggested the predictable plug-in $(\lambda_t^{\mathrm{PrPl}})_{t=1}^\infty$ given by (15), where the additional $\log(t+1)$ is included in an attempt to enforce $W_t = \widetilde{O}(\sqrt{\log t/t})$.

The above calculations are only used as guiding principles to sharpen the confidence sets, but *all* such schemes retain the validity guarantee. As long as $(\lambda_t)_{t=1}^\infty$ is $[0,1)$-valued and predictable, we have that $(M_t^E(\mu))_{t=0}^\infty$ is a test supermartingale for $\mathcal{P}^\mu$ which can be used in Theorem 1 to obtain different valid CSs for μ.

Foreshadowing our attempt to generalize this procedure in the next section, notice that the exponential function was used throughout to ensure nonnegativity, but that any other test supermartingale would have sufficed. In fact, if a martingale is used in place of a supermartingale, then Ville's inequality is tighter.

Next, we present a test *martingale*, removing a source of looseness in the confidence sets derived thus far. We discuss its betting interpretation, provide other guiding principles for setting λ_i (equivalently, for betting), which will involve attempting to maximize the expected log-wealth in the betting game.

4. The capital process, betting, and martingales

In Section 3, we generalized the Cramer-Chernoff method to derive predictable plug-in exponential supermartingales and used this result to obtain tight empirical Bernstein CSs and CIs. In this section, we consider an alternative process which can be interpreted as the wealth accumulated from a series of bets in a game. This process is a central object of study in the game-theoretic probability literature where it is referred to as the *capital process* [Shafer and Vovk, 2001]. We discuss its connections to the purely statistical goal of constructing CSs and CIs and demonstrate how these sets improve on Cramer-Chernoff approaches, including the empirical Bernstein confidence sets of the previous section.

Consider the same setup as in Section 3: we observe an infinite sequence of conditionally mean-μ random variables, $(X_t)_{t=1}^\infty \sim P$ from some distribution $P \in \mathcal{P}^\mu$. Define the *capital process* $\mathcal{K}_t(m)$ for any $m \in [0,1]$,

$$\mathcal{K}_t(m) := \prod_{i=1}^t (1 + \lambda_i(m) \cdot (X_i - m)), \tag{23}$$

with $\mathcal{K}_0(m) := 1$ and where $(\lambda_t(m))_{t=1}^{\infty}$ is a $(-1/(1-m), 1/m)$-valued predictable sequence, and thus $\lambda_t(m)$ can depend on X_1^{t-1}. Note that for each $t \geq 1$, we have $X_t \in [0,1]$, $m \in [0,1]$ and $\lambda_t(m) \in (-1/(1-m), 1/m)$. Here and below, $1/m$ should be interpreted as ∞ when $m = 0$ and similarly for $1/(1-m)$ and $m = 1$, respectively. Importantly, $(1 + \lambda_t(m) \cdot (X_t - m)) \in [0, \infty)$, and thus $\mathcal{K}_t(m) \geq 0$ for all $t \geq 1$. Following similar techniques to the previous section, the reader may easily check that $\mathcal{K}_t(\mu)$ is a test martingale. Moreover, we have the stronger result summarized in the following central proposition.

PROPOSITION 2. *Suppose a draw from some distribution P yields a sequence $X_1, X_2, \dots$ of $[0,1]$-valued random variables, and let $\mu \in [0,1]$ be a constant. The following four statements imply each other:*

(a) $\mathbb{E}_P(X_t \mid \mathcal{F}_{t-1}) = \mu$ *for all* $t \in \mathbb{N}$, *where* $\mathcal{F}_{t-1} = \sigma(X_1, \dots, X_{t-1})$.

(b) There exists a constant $\lambda \in \mathbb{R}\backslash\{0\}$ *for which* $(\mathcal{K}_t(\mu))_{t=0}^{\infty}$ *is a strictly positive test martingale for* P.

(c) For every fixed $\lambda \in (-\frac{1}{1-\mu}, \frac{1}{\mu})$, $(\mathcal{K}_t(\mu))_{t=0}^{\infty}$ *is a test martingale for* P.

(d) For every $(-\frac{1}{1-\mu}, \frac{1}{\mu})$*-valued predictable sequence* $(\lambda_t)_{t=1}^{\infty}$, $(\mathcal{K}_t(\mu))_{t=0}^{\infty}$ *is a test martingale for* P.

Further, the intervals $(-\frac{1}{1-\mu}, \frac{1}{\mu})$ *mentioned above can be replaced by any subinterval containing at least one nonzero value, like* $[-1,1]$ *or* $(-0.5, 0.5)$. *Finally, every test martingale for* $\mathcal{P}^\mu$ *is of the form* $(\mathcal{K}_t(\mu))$ *for some predictable sequence* (λ_t).

The proof can be found in Section A.3. While the subsequent theorems will primarily make use of $(a) \implies (d)$, the above proposition establishes a core fact: the assumption of the (conditional) means being identically μ is an *equivalent restatement* of our capital process being a test martingale. Thus, test martingales are not simply "technical tools" to deal with means of bounded random variables, they are fundamentally at the very heart of the problem definition itself.

Proposition 2 can be generalized to another remarkable, yet simple, result: for any set of distributions $\mathcal{S}$, *every* test martingale for $\mathcal{S}$ has the same form.

PROPOSITION 3 (UNIVERSAL REPRESENTATION). *For any arbitrary set of (possibly unbounded) distributions* $\mathcal{S}$, (M_t) *is a test martingale for* $\mathcal{S}$ *if and only if* $M_t = \prod_{i=1}^{t}(1 + \lambda_i Z_i)$ *for some* $Z_i \geq -1$ *such that* $\mathbb{E}_S[Z_i | \mathcal{F}_{i-1}] = 0$ *for every* $S \in \mathcal{S}$, *and some predictable* λ_i *such that* $\lambda_i Z_i \geq -1$. *The same claim also holds for test supermartingales for* $\mathcal{S}$, *with the aforementioned "*$= 0$*" replaced by "*≤ 0*".*

The proof can be found in Section A.4. The above proposition immediately makes this paper's techniques actionable for a wide class of nonparametric testing and estimation problems. We give an example relating to quantiles later.

4.1. Connections to betting

It is worth pausing to clarify how the capital process $\mathcal{K}_t(m)$ and Proposition 2 can be viewed in terms of betting. We imagine that nature implicitly posits a hypothesis H_0^m — which we treat as a game providing us a chance to make money if the hypothesis is wrong, by repeatedly betting some of our capital against H_0^m. We start the game with a capital of 1 (i.e. $\mathcal{K}_0(m) := 1$), and design a bet of $b_t := s_t|\lambda_t^m|$ at each step, where $s_t \in \{-1, 1\}$. Setting $s_t := 1$ indicates that we believe that $\mu > m$ while $s_t := -1$ indicates the opposite. $|\lambda_t^m|$ indicates the amount of our capital that we are willing to put at stake at time t: setting $\lambda_t^m = 0$ results in neither losing nor gaining any capital regardless of the outcome, while setting $\lambda_t^m \in \{-1/(1-m), 1/m\}$ means that we are willing to risk all of our capital on the next outcome.

However, if H_0^m is true (i.e. $m = \mu$), then by Proposition 2, our capital process is a martingale. In betting terms, no matter how clever a betting strategy $(\lambda_t^m)_{t=1}^\infty$ we devise, we cannot expect to make (or lose) money at each step. If on the other hand, H_0^m is false, then a clever betting strategy will make us a lot of money. In statistical terms, when our capital exceeds $1/\alpha$, we can confidently reject the hypothesis H_0^m since if it were true (and the game were fair) then by Ville's inequality [Ville, 1939], the a priori probability of this *ever* occurring is at most α. We imagine simultaneously playing this game with $H_0^{m'}$ for each $m' \in [0, 1]$. At any time t, the games $m' \in [0, 1]$ for which our capital is small ($< 1/\alpha$) form a CS.

Both the Cramer-Chernoff processes of Section 3 and $\mathcal{K}_t(m)$ are nonnegative and tend to increase when $\mu > m$. However, only $\mathcal{K}_t(m)$ is a *test martingale* when $m = \mu$; the others are test supermartingales. A test martingale is the wealth accumulated in a "fair game" where our capital stays constant in expectation, while a test supermartingale is the wealth accumulated in a game where our capital is expected to decrease (not strictly). Larger values of capital correspond to rejecting H_0^m more readily. Therefore, test supermartingales tend to yield conservative tests compared to their martingale counterparts.

More generally, *every* nonnegative supermartingale can be regarded as the wealth process of a gambler playing a game with odds that are fair or stacked against them. In other words, there is a one-to-one correspondence between wealths of hypothetical gamblers and nonnegative supermartingales. Taking this perspective, every statement involving nonnegative supermartingales (and thus likelihood ratios) are statements about betting, and vice versa. Mixture methods that combine nonnegative supermartingales are simply strategies to hedge across various instruments available to the gambler. Thus, the gambling analogy can be entirely dropped, and our results would find themselves comfortably nestled in the rich literature on martingale methods for concentration inequalities, but we mention the betting analogy for intuition so that the mathematics are animated and easier to absorb.

Ville introduced martingales into modern mathematical probability theory, and centered them around their betting interpretation. Since then, ideas from betting have appeared in various fields, including probability theory, statistical testing and estimation, information theory, and online learning theory. While our paper focuses on the utility of betting in some statistical inference tasks, Section F provides a brief overview of the use of betting in other mathematical disciplines.

4.2. Connections to likelihood ratios

As alluded to in the previous subsection, useful intuition is provided via the connection to likelihood ratios. $\mathcal{K}_t(m)$ is a "composite" test martingale for $\mathcal{P}^m$, meaning that it is a nonnegative martingale starting at one for every $P \in \mathcal{P}^m$ (recall that P is a distribution over infinite sequences of observations with conditional mean m).

If we were dealing with a single distribution such as Q^∞, meaning a product distribution where every observation is drawn iid from Q, then one may pick any alternative Q' that is absolutely continuous with respect to Q, to observe that the likelihood ratio $\prod_{i=1}^t Q'(X_i)/Q(X_i)$ is a test martingale for Q^∞.

However, since $\mathcal{P}^m$ is highly composite and nonparametric and is not even dominated by a single measure (as it contains atomic measures, continuous measures, and all their mixtures), it is unclear how one can even begin to write down a likelihood ratio. Nevertheless, Ramdas et al. [2020, Proposition 4] show that if (M_t) is a composite test martingale for any $\mathcal{S}$, then for every distribution $Q \in \mathcal{S}$, M_t equals the likelihood ratio of some Q' against Q (where Q' depends on Q).

Thus, not only is every likelihood ratio a test martingale, but every (composite) test martingale can also be represented as a likelihood ratio. Hence, in a formal sense, test martingales are nonparametric composite generalizations of likelihood ratios, which are at the very heart of statistical inference. When this observation is combined with Proposition 2, it should be no surprise any longer that the capital process $\mathcal{K}_t(m)$ (even devoid of any betting interpretation) is fundamental to the problem at hand. In Section E.6 we also observe connections to the empirical likelihood of Owen [2001] and the dual likelihood of [Mykland, 1995].

4.3. Adaptive, constrained adversaries

Despite the analogies to betting, the game described so far appears to be purely stochastic in the sense that nature simply commits to a distribution $P \in \mathcal{P}^\mu$ for some unknown $\mu \in [0, 1]$ and presents us observations from P. However, Proposition 2 can be extended to a more adversarial setup, but with a constrained adversary.

To elaborate, recall the difference between $\mathcal{Q}$ and $\mathcal{P}$ from the start of Section 2 and consider a game with three players: an adversary, nature, and the statistician. First, the adversary commits to a $\mu \in [0, 1]$. Then, the game proceeds in rounds. At the start of round t, the statistician publicly discloses the bets for every m, which could depend on $X_1, \dots, X_{t-1}$. The adversary picks a distribution $Q_t \in \mathcal{Q}^\mu$, which could depend on $X_1, \dots, X_{t-1}$ and the statistician's disclosed bets, and hands Q_t to nature. Nature simply acts like an arbitrator, first verifying that the adversary chose a Q_t with mean μ, and then draws $X_t \sim Q_t$ and presents X_t to the statistician.

In this fashion, the adversary does not need to pick μ and $P \in \mathcal{P}^\mu$ at the start of the interaction, which is the usual stochastic setup, but can instead build the distribution P in a data-dependent fashion over time. In other words, the adversary does not commit to a distribution P, but instead to a *rule for building* P from the data. Of course, they do not need to disclose this rule, or even be able express what this rule would do on any other hypothetical outcomes other than the one observed. The results in this paper, which build on the central Proposition 2, continue to hold in this more general interaction model.

A geometric reason why we can move from the stochastic model first described to the above (constrained) adversarial model, is that the above distribution P lies in the "fork convex hull" of $\mathcal{P}^{\mu}$. Fork-convexity is a sequential analogue of convexity [Ramdas et al., 2021]. Informally, the fork-convex hull of a set of distributions over sequences is the set of predictable plug-ins of these distributions, and is much larger than their convex hull (mixtures). If a process is a nonnegative martingale under every distribution in a set, then it is also a nonnegative martingale under every distribution in the fork convex hull of that set. No results about fork convexity are used anywhere in this paper, and we only mention it for the mathematically curious.

4.4. *The hedged capital process*

We now return to the purely statistical problem of using the capital process $\mathcal{K}_t(m)$ to construct time-uniform CSs and fixed-time CIs. We might be tempted to use $\mathcal{K}_t(\mu)$ as the nonnegative martingale in Theorem 1 to conclude that $\mathfrak{B}_t := \{m \in [0,1] : \mathcal{K}_t(m) < 1/\alpha\}$ forms a $(1-\alpha)$-CS for μ. Unlike the empirical Bernstein CS of Section 3, $\mathfrak{B}_t$ cannot be computed in closed-form. Instead, we theoretically need to compute the family of processes $\{\mathcal{K}_t(m)\}_{m\in[0,1]}$ and include those $m \in [0,1]$ for which $\mathcal{K}_t(m)$ remains below $1/\alpha$. This is not practical as the parameter space $[0,1]$ is uncountably infinite. But if we know a priori that $\mathfrak{B}_t$ is guaranteed to produce an interval for each t, then it is straightforward to find a superset of $\mathfrak{B}_t$ by either performing a grid search on $(0, 1/g, 2/g, \dots, (g-1)/g, 1)$ for some large $g \in \mathbb{N}$, or by employing root-finding algorithms. This motivates the *hedged capital process*, defined for any $\theta, m \in [0,1]$ as

$$\mathcal{K}_t^{\pm}(m) := \max\left\{\theta \mathcal{K}_t^{+}(m), (1-\theta)\mathcal{K}_t^{-}(m)\right\}, \tag{24}$$

$$\text{where} \quad \mathcal{K}_t^{+}(m) := \prod_{i=1}^{t}(1 + \lambda_i^{+}(m)\cdot(X_i - m)),$$

$$\text{and} \quad \mathcal{K}_t^{-}(m) := \prod_{i=1}^{t}(1 - \lambda_i^{-}(m)\cdot(X_i - m)),$$

and $(\lambda_t^{+}(m))_{t=1}^{\infty}$ and $(\lambda_t^{-}(m))_{t=1}^{\infty}$ are predictable sequences of $[0, \frac{1}{m})$- and $[0, \frac{1}{1-m})$-valued random variables, respectively.

$\mathcal{K}_t^{\pm}(m)$ can be viewed from the betting perspective as dividing one's capital into proportions of θ and $(1-\theta)$ and making two series of simultaneous bets, positing that $\mu \geq m$, and $\mu < m$, respectively which accumulate capital in $\mathcal{K}_t^{+}(m)$ and $\mathcal{K}_t^{-}(m)$. If $\mu \neq m$, then we expect that one of these strategies will perform poorly, while we expect the other to make money in the long term. If $\mu = m$, then we expect neither strategy to make money. The maximum of these processes is upper-bounded by their convex combination,

$$\mathcal{M}_t^{\pm} := \theta\mathcal{K}_t^{+} + (1-\theta)\mathcal{K}_t^{-}.$$

Both $\mathcal{K}_t^{\pm}$ and $\mathcal{M}_t^{\pm}$ can be used for Step (b) of Theorem 1 to yield a CS. Empirically, both yield intervals, but only the former provably so.

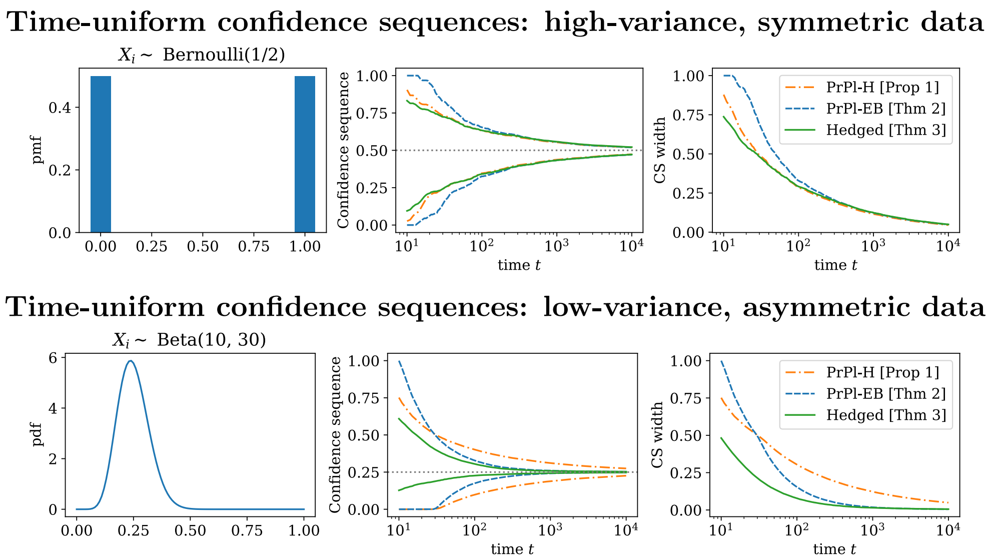

Figure 4. Predictable plug-in Hoeffding, empirical Bernstein, and hedged capital CSs under two distributional scenarios. Notice that the latter roughly matches the others in the Bernoulli$(1/2)$ case, but shines in the low-variance, asymmetric scenario.

THEOREM 3 (HEDGED CAPITAL CS [**HEDGED**]). *Suppose* $(X_t)_{t=1}^\infty \sim P$ *for some* $P \in \mathcal{P}^\mu$. *Let* $(\tilde\lambda_t^+)_{t=1}^\infty$ *and* $(\tilde\lambda_t^-)_{t=1}^\infty$ *be real-valued predictable sequences not depending on* m, *and for each* $t \geq 1$ *let*

$$\lambda_t^+(m) := |\tilde\lambda_t^+| \wedge \frac{c}{m}, \quad \lambda_t^-(m) := |\tilde\lambda_t^-| \wedge \frac{c}{1-m}, \tag{25}$$

for some $c \in [0,1)$ *(some reasonable defaults being* $c = 1/2$ *or* $3/4$*). Then*

$$\mathfrak{B}_t^\pm := \left\{ m \in [0,1] : \mathcal{K}_t^\pm(m) < 1/\alpha \right\} \quad \text{forms a } (1-\alpha)\text{-CS for } \mu,$$

as does its running intersection $\bigcap_{i \leq t} \mathfrak{B}_i^\pm$. *Further,* $\mathfrak{B}_t^\pm$ *is an interval for each* $t \geq 1$. *Finally, replacing* $\mathcal{K}_t^\pm(m)$ *by* $\mathcal{M}_t^\pm(m)$ *yields a tighter* $(1-\alpha)$*-CS for* μ.

For reasons given in Section B.1, we recommend setting $\tilde\lambda_t^+ = \tilde\lambda_t^- = \lambda_t^{\mathrm{PrPl}\pm}$ as

$$\lambda_t^{\mathrm{PrPl}\pm} := \sqrt{\frac{2\log(2/\alpha)}{\widehat\sigma_{t-1}^2 t \log(t+1)}}, \quad \widehat\sigma_t^2 := \frac{1/4 + \sum_{i=1}^t (X_i - \widehat\mu_i)^2}{t+1}, \quad \text{and} \quad \widehat\mu_t := \frac{1/2 + \sum_{i=1}^t X_i}{t+1}, \tag{26}$$

for each $t \geq 1$, and truncation level $c := 1/2$ or $3/4$; see Figure 4. A reasonable point estimator for μ is $\operatorname{argmin}_{m \in [0,1]} \mathcal{K}_t^\pm(m)$ or $\operatorname{argmin}_{m \in [0,1]} \mathcal{M}_t^\pm(m)$ (see Figure 18).

REMARK 2. *Since* $\mathcal{K}_t^\pm(m) \leq \mathcal{M}_t^\pm(m)$, *the latter confidence sequence is tighter. In the proof of Theorem 3, we use a property of the* max *function to establish quasiconvexity of* $\mathcal{K}_t^\pm(m)$, *implying that* $\mathfrak{B}_t^\pm$ *is an interval. We find the difference in empirical performance negligible (Figure 5). For the interested reader, Section E.4 constructs a (pathological) CS that is almost surely not an interval.*

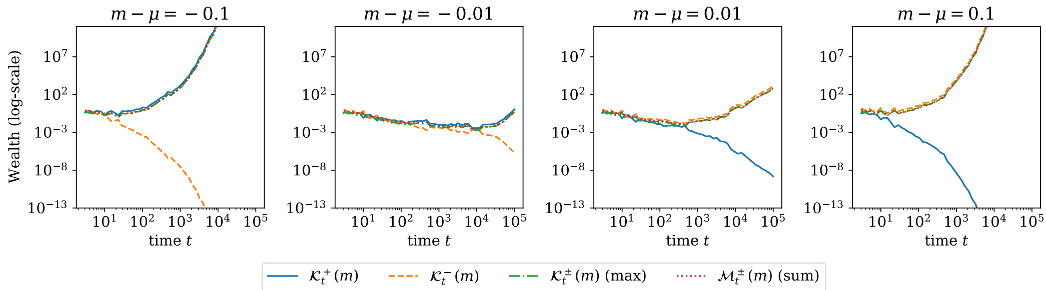

Figure 5. A comparison of capital processes $\mathcal{K}_t^+(m)$, $\mathcal{K}_t^-(m)$, the hedged capital process $\mathcal{K}_t^\pm(m)$, and its upper-bounding nonnegative martingale, $\mathcal{M}_t^\pm(m)$ under four alternatives (from left to right): $m \ll \mu$, $m < \mu$, $m > \mu$, $m \gg \mu$. When $m < \mu$, we see that $\mathcal{K}_t^+(m)$ increases, while $\mathcal{K}_t^-(m)$ approaches zero, but the opposite is true when $m > \mu$. Notice that not much is gained by taking a sum $\mathcal{M}_t^\pm(m)$ rather than a maximum $\mathcal{K}_t^\pm(m)$, since one of $\mathcal{K}_t^+(m)$ and $\mathcal{K}_t^-(m)$ vastly dominates the other, depending on whether $m > \mu$ or $m < \mu$.

REMARK 3. *Theorem 3 yields tight hedged CIs for a fixed sample size n. Recalling (26), we recommend using $\bigcap_{i \leq n} \mathfrak{B}_i^\pm$, and setting $\tilde{\lambda}_t^+ = \tilde{\lambda}_t^- = \tilde{\lambda}_t^\pm$ given by*

$$\tilde{\lambda}_t^\pm := \sqrt{\frac{2\log(2/\alpha)}{n\widehat{\sigma}_{t-1}^2}}. \tag{27}$$

We refer to the resulting CI as the "hedged capital confidence interval" or **[Hedged-CI]** *for short, and demonstrate its superiority to past work in Figure 6.*

Similar to the discussion after Remark 1, if $X_1, \dots, X_n$ are independent, then one can permute the data many times and average the resulting capital processes to effectively derandomize the procedure.

The proof of Theorem 3 is in Section A.5. Unlike the empirical Bernstein-type CSs and CIs of Section 3, those based on the hedged capital process are not necessarily symmetric. In fact, we empirically find through simulations that these CSs and CIs are able to adapt and benefit from this asymmetry (see Figures 4 and 6). While it is not obvious from the definition of $\mathfrak{B}_t^\pm$, bets can be chosen such that hedged capital CSs and CIs converge at the optimal rates of $O(\sqrt{\log\log t/t})$ and $O(1/\sqrt{n})$, respectively (see Section E.2) and such that for sufficiently large n, hedged capital CIs almost surely dominate those based on Hoeffding's inequality (see Section E.1). However, the implications of time-uniform convergence rates are subtle, and optimal rates are not always desirable in practical applications (see [Howard et al., 2021, Section 3.5]). Nevertheless, we find that hedged capital CSs and CIs significantly outperform past works even for small sample sizes (see Section C). Some additional tools for visualizing CSs across α and t are provided in Section D.5.

In Section B, we discuss some guiding principles for deriving powerful betting strategies, presenting the hedged capital CSs and CIs as special cases along with the following game-theoretic betting schemes:

- Growth rate adaptive to the particular alternative (GRAPA),

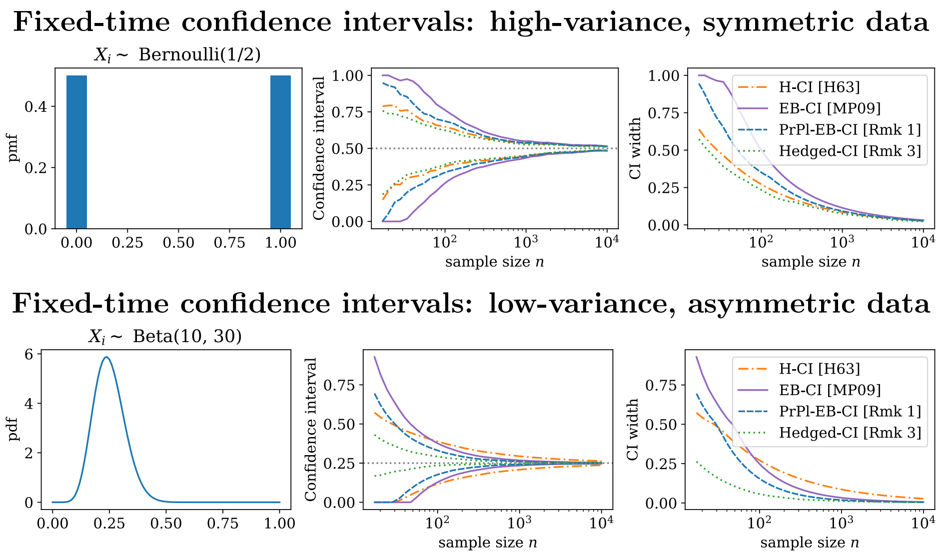

Figure 6. Hoeffding (H), empirical Bernstein (EB), and hedged capital CIs under two distributional scenarios. Similar to the time-uniform setting, the betting approach tends to outperform the other bounds, especially for low-variance, asymmetric data.

- Approximate GRAPA (aGRAPA),
- Lower-bound on the wealth (LBOW),
- Online Newton step-m (ONS-m),
- Diversified Kelly betting (dKelly),
- Confidence boundary bets (ConBo), and
- Sequentially rebalanced portfolio (SRP).

Each of these betting strategies have their respective benefits, whether computational, conceptual, or statistical which are discussed further in Section B.

5. Betting while sampling without replacement (WoR)

This section tackles a slightly different problem, that of sampling without replacement (WoR) from a finite set of real numbers in order to estimate its mean. Importantly, the N numbers in the finite population $(x_1, \ldots, x_N)$ are fixed and nonrandom. What is random is only the order of observation; the model for sampling uniformly at random without replacement (WoR) posits that at time $t \geq 1$,

$$X_t \mid (X_1, \ldots, X_{t-1}) \sim \text{Uniform}\left((x_1, \ldots, x_N) \backslash (X_1, \ldots, X_{t-1})\right). \tag{28}$$

All probabilities are thus to be understood as solely arising from observing fixed entities in a random order, with no distributional assumptions being made on the finite population. We consider the same canonical filtration $\mathcal{F} = (\mathcal{F}_t)_{t=0}^N$ as before.

For $t \geq 1$, let $\mathcal{F}_t := \sigma(X_1^t)$ be the sigma-field generated by $X_1, \ldots, X_t$ and let $\mathcal{F}_0$ be the empty sigma-field. For succinctness, we use the notation $[a] := \{1, \ldots, a\}$.

For each $m \in [0,1]$, let $\mathcal{L}^m := \{x_1^N \in [0,1]^N : \sum_{i=1}^N x_i/N = m\}$ be the set of all unordered lists of $N \geq 2$ real numbers in $[0,1]$ whose average is m. For instance, $\mathcal{L}^0$ and $\mathcal{L}^1$ are both singletons, but otherwise $\mathcal{L}^m$ is uncountably infinite. Let $\mathcal{P}^m$ be the set of all measures on $\mathcal{F}_N$ that are formed as follows: pick an arbitrary element of $\mathcal{L}^m$, apply a uniformly random permutation, and reveal the elements one by one. Thus, every element of $\mathcal{P}^m$ is a uniform measure on the $N!$ permutations of some element in $\mathcal{L}^m$, so there is a one-to-one mapping between $\mathcal{L}^m$ and $\mathcal{P}^m$.

Define $\mathcal{P} := \bigcup_m \mathcal{P}^m$ and let μ represent the true unknown mean, meaning that the data is drawn from some $P \in \mathcal{P}^\mu$. For every $m \in [0,1]$, we posit a composite null hypothesis $H_m^0 : P \in \mathcal{P}^m$, but clearly only one of these nulls is true. We will design betting strategies to test these nulls and thus find efficient confidence intervals or sequences for μ. It is easier to present the sequential case first, since that is arguably more natural for sampling WoR, and discuss the fixed-time case later.

5.1. Existing (super)martingale-based confidence sequences or tests

Several papers have considered estimating the mean of a finite set of nonrandom numbers when sampling WoR, often by constructing concentration inequalities [Hoeffding, 1963, Serfling, 1974, Bardenet and Maillard, 2015, Waudby-Smith and Ramdas, 2020]. Notably, Hoeffding [1963] showed that the same bound for sampling with replacement (2) can be used when sampling WoR. Serfling [1974] improved on this bound, which was then further refined by Bardenet and Maillard [2015]. While test supermartingales appeared in some of the aforementioned works, Waudby-Smith and Ramdas [2020] identified better test supermartingales which yield explicit Hoeffding- and empirical Bernstein-type concentration inequalities and CSs for sampling WoR that significantly improved on previous bounds. Consider their exponential Hoeffding-type supermartingale,

$$M_t^{\text{H-WoR}} := \exp\left\{\sum_{i=1}^t \left[\lambda_i \left(X_i - \mu + \frac{1}{N-(i-1)}\sum_{j=1}^{i-1}(X_j - \mu)\right) - \psi_H(\lambda_i)\right]\right\}, \tag{29}$$

and their exponential empirical Bernstein-type supermartingale,

$$M_t^{\text{EB-WoR}} := \exp\left\{\sum_{i=1}^t \left[\lambda_i \left(X_i - \mu + \frac{1}{N-(i-1)}\sum_{j=1}^{i-1}(X_j - \mu)\right) - v_i\psi_E(\lambda_i)\right]\right\}, \tag{30}$$

where $(\lambda_t)_{t=1}^N$ is any predictable λ-sequence (real-valued for $M_t^{\text{H-WoR}}$, but $[0,1)$-valued for $M_t^{\text{EB-WoR}}$), $v_i = 4(X_i - \widehat{\mu}_{i-1})^2$ as before, and $\psi_H(\cdot)$ and $\psi_E(\cdot)$ are defined as in Section 3. Defining $M_0^{\text{H-WoR}} \equiv M_0^{\text{EB-WoR}} := 1$, Waudby-Smith and Ramdas [2020] prove that $(M_t^{\text{H-WoR}})_{t=0}^N$ and $(M_t^{\text{EB-WoR}})_{t=0}^N$ are test supermartingales with respect to $\mathcal{F}$, and hence can be used in Step (b) of Theorem 1.

In recent work on election audits, Stark [2020] credits Harold Kaplan for proposing

$$M_t^K := \int_0^1 \prod_{i=1}^t \left(1 + \gamma\left[X_i \frac{1-(i-1)/N}{\mu - \sum_{j=1}^{i-1} X_i/N} - 1\right]\right) d\gamma. \tag{31}$$

The "Kaplan martingale" $(M_t^K)_{t=0}^N$ was employed for election auditing, but it is a polynomial of degree t and is computationally expensive for large t [Stark, 2020].

Next, we mimic the approach of Section 4 to derive a capital process for sampling WoR. We then derive WoR analogues of the efficient betting strategies from Section B.

5.2. The capital process for sampling without replacement

Define the predictable sequence $(\mu_t^{\mathrm{WoR}})_{t\in[N]}$ where

$$\mu_t^{\mathrm{WoR}} := \mathbb{E}[X_t | \mathcal{F}_{t-1}] = \frac{N\mu - \sum_{i=1}^{t-1} X_i}{N-(t-1)}. \tag{32}$$

It is clear that $\mu_t^{\mathrm{WoR}} \in [0,1]$, since it is the mean of the unobserved elements of $\{x_i\}_{i\in[N]}$. $(\mu_t^{\mathrm{WoR}})_{t\in[N]}$ is unobserved since μ is unknown, so it is helpful to define

$$m_t^{\mathrm{WoR}} := \frac{Nm - \sum_{i=1}^{t-1} X_i}{N-(t-1)}. \tag{33}$$

Now, let $(\lambda_t(m))_{t=1}^N$ be a predictable sequence such that $\lambda_t(m)$ is $\left(-\frac{1}{1-m_t^{\mathrm{WoR}}}, \frac{1}{m_t^{\mathrm{WoR}}}\right)$-valued. Define the *without-replacement capital process* $\mathcal{K}_t^{\mathrm{WoR}}(m)$,

$$\mathcal{K}_t^{\mathrm{WoR}}(m) := \prod_{i=1}^t \left(1 + \lambda_i(m)\cdot(X_i - m_i^{\mathrm{WoR}})\right) \tag{34}$$

with $\mathcal{K}_0^{\mathrm{WoR}}(m) := 1$. The following result is analogous to Proposition 2.

Proposition 4. *Let X_1^N be a WoR sample from $x_1^N \in [0,1]^N$. The following two statements imply each other:*

(a) $\mathbb{E}_P\left(X_t \mid \mathcal{F}_{t-1}\right) = \mu_t^{\mathrm{WoR}}$ for each $t \in [N]$.

(b) For every predictable sequence with $\lambda_t(m) \in \left(-\frac{1}{(1-\mu_t^{\mathrm{WoR}})}, \frac{1}{\mu_t^{\mathrm{WoR}}}\right)$, $(\mathcal{K}_t^{\mathrm{WoR}}(\mu))_{t=0}^{\infty}$ is a test martingale.

The other claims within Proposition 2 also hold above with minor modification, but we do not mention them again for brevity. Further, Proposition 3 technically covers WoR sampling as well. We now present a "hedged" capital process and powerful betting schemes for sampling WoR, to construct a CS for $\mu = \frac{1}{N}\sum_{i=1}^N x_i$.

5.3. Powerful betting schemes

Similar to Section 4.4, define the hedged capital process for sampling WoR:

$$\mathcal{K}_t^{\mathrm{WoR},\pm}(m) := \max\left\{\theta \prod_{i=1}^t \left(1 + \lambda_i^+(m)\cdot(X_i - m_t^{\mathrm{WoR}})\right),\right.$$
$$\left.(1-\theta)\prod_{i=1}^t\left(1 - \lambda_i^-(m)\cdot(X_i - m_t^{\mathrm{WoR}})\right)\right\}$$

for some predictable $(\lambda_t^+(m))_{t=1}^N$ and $(\lambda_t^-(m))_{t=1}^N$ taking values in $[0, 1/m_t^{\text{WoR}}]$ and $[0, 1/(1-m_t^{\text{WoR}})]$ at time t, respectively. Using $\left(\mathcal{K}_t^{\text{WoR},\pm}(m)\right)_{t=0}^{\infty}$ as the process in Step (b) of Theorem 1, we obtain the CS summarized in the following theorem.

THEOREM 4 (WOR HEDGED CAPITAL CS **[HEDGED-WOR]**). *Given a finite population $x_1^N \in [0,1]^N$ with mean $\mu := \frac{1}{N}\sum_{i=1}^N x_i = \mu$, suppose that $X_1, X_2, \dots X_N$ are sampled WoR from x_1^N. Let $(\dot\lambda_t^+)_{t=1}^\infty$ and $(\dot\lambda_t^-)_{t=1}^\infty$ be real-valued predictable sequences not depending on m, and for each $t \geq 1$ let*

$$\lambda_t^+(m) := |\dot\lambda_t^+| \wedge \frac{c}{m_t^{\text{WoR}}}, \quad \lambda_t^-(m) := |\dot\lambda_t^-| \wedge \frac{c}{1-m_t^{\text{WoR}}},$$

for some $c \in [0,1)$ (some reasonable defaults being $c = 1/2$ or $3/4$). Then

$$\mathfrak{B}_t^{\pm,\text{WoR}} := \left\{ m \in [0,1] : \mathcal{K}_t^{\pm,\text{WoR}}(m) < 1/\alpha \right\} \quad \textit{forms a } (1-\alpha)\textit{-CS for } \mu,$$

as does $\bigcap_{i\leq t} \mathfrak{B}_i^{\pm,\text{WoR}}$. Furthermore, $\mathfrak{B}_t^{\pm,\text{WoR}}$ is an interval for each $t \geq 1$.

The proof of Theorem 4 is in Section A.9. We recommend setting $\dot\lambda_t^+ = \dot\lambda_t^- = \lambda_t^{\text{PrPl}\pm}$ as was done earlier in (26); for each $t \geq 1$, and $c := 1/2$, let

$$\lambda_t^{\text{PrPl}\pm} := \sqrt{\frac{2\log(2/\alpha)}{\widehat\sigma_{t-1}^2 t \log(t+1)}}, \quad \widehat\sigma_t^2 := \frac{\frac{1}{4} + \sum_{i=1}^t (X_i - \widehat\mu_i)^2}{t+1}, \text{ and } \widehat\mu_t := \frac{\frac{1}{2} + \sum_{i=1}^t X_i}{t+1},$$

See Figure 7 for a comparison to the best prior work.

REMARK 4. *As before, we can use Theorem 4 to derive powerful CIs for the mean of a nonrandom set of bounded numbers given a fixed sample size $n \leq N$. We recommend using $\bigcap_{i\leq n} \mathfrak{B}_i^{\pm,\text{WoR}}$, and setting $\dot\lambda_t^+ = \dot\lambda_t^- = \dot\lambda_t^\pm$ as in (27): $\dot\lambda_t^\pm := \sqrt{\frac{2\log(2/\alpha)}{n\widehat\sigma_{t-1}^2}}$. We refer to the resulting CI as **[Hedged-WoR-CI]**; see Figure 8.*

Notice that constructing a WoR test martingale only relies on changing the fixed conditional mean μ to the time-varying conditional mean $\mu_t^{\text{WoR}} := \frac{N\mu - \sum_{i=1}^{t-1} X_i}{N-t+1}$ and now designing $(-1/(1-\mu_t^{\text{WoR}}), 1/\mu_t^{\text{WoR}})$-valued bets instead of $(-1/(1-\mu), 1/\mu)$-valued ones. In this way, it is possible to adapt any of the betting strategies in Section B to sampling WoR, yielding a wide array of solutions to this estimation problem.

5.4. Relationship to composite null testing

This paper focuses primarily on estimation, but we end with a note that our CSs (or CIs) yield valid, sequential (or batch) tests for composite null hypotheses $H_0 : \mu \in S$ for any $S \subset [0,1]$. Specifically, for any of our capital processes $\mathcal{K}_t(m)$,

$$\mathfrak{p}_t := \sup_{m\in S} \frac{1}{\mathcal{K}_t(m)}$$

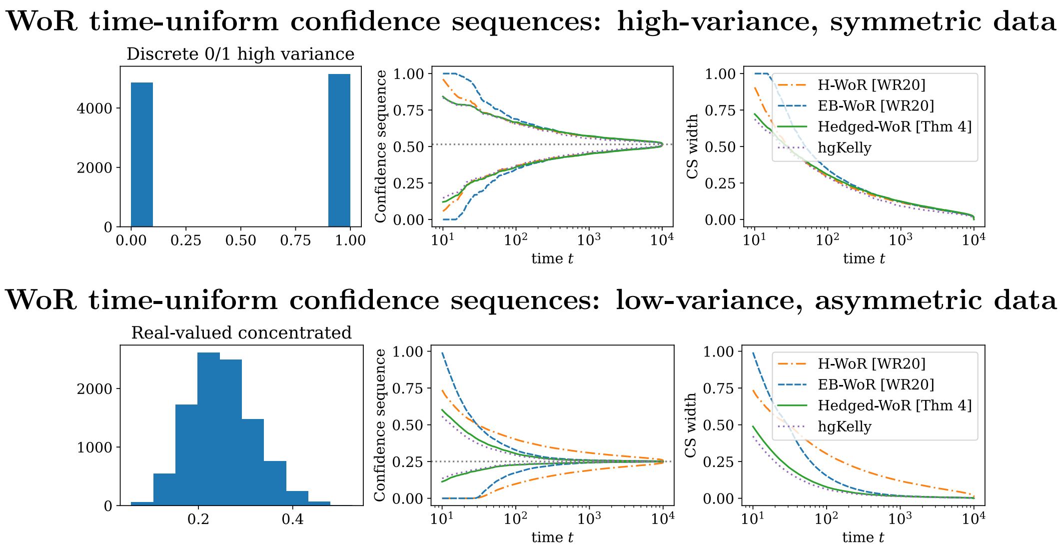

Figure 7. Without-replacement betting CSs versus the predictable plug-in supermartingale-based CSs [Waudby-Smith and Ramdas, 2020]. Similar to the with-replacement case, the betting approach matches or vastly outperforms past state-of-the art methods.

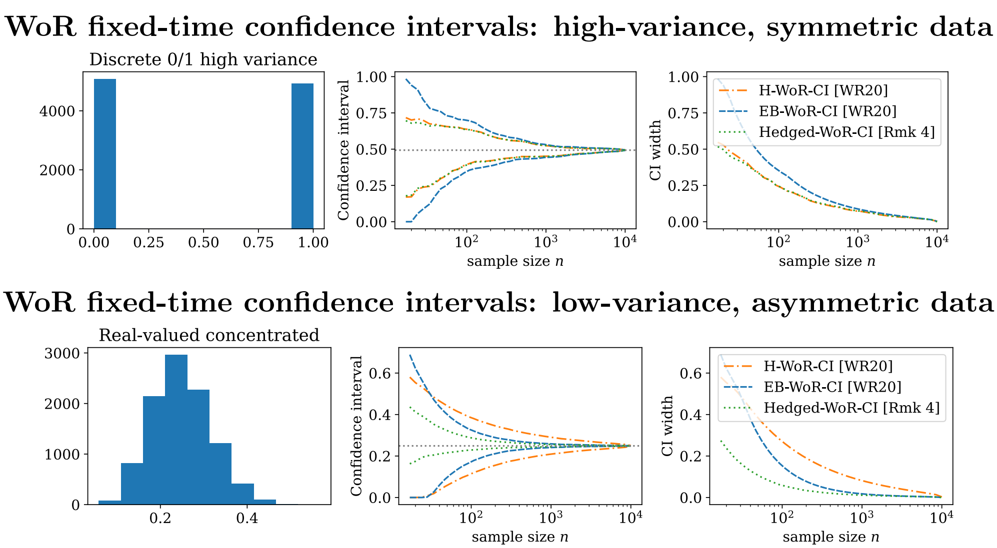

Figure 8. WoR analogue of the hedged capital CI versus the WoR Hoeffding- and empirical Bernstein-type CIs [Waudby-Smith and Ramdas, 2020]. Similar to with-replacement, the betting approach has excellent performance.

is an "anytime-valid p-value" for H_0, as is $\widetilde{\mathfrak{p}}_t := \inf_{s \leq t} \mathfrak{p}_s$, meaning that

$$\sup_{P \in \bigcup_{m \in S} \mathcal{P}^m} P(\mathfrak{p}_\tau \leq \alpha) \leq \alpha \text{ for arbitrary stopping times } \tau.$$

Alternately, $\mathfrak{p}_t$ is also the smallest α for which our $(1-\alpha)$-CS does not intersect S. Similarly, $\mathfrak{e}_t := \inf_{m \in S} \mathcal{K}_t(m)$ is an "e-process" for H_0, meaning that

$$\sup_{P \in \bigcup_{m \in S} \mathcal{P}^m} \mathbb{E}_P[\mathfrak{e}_\tau] \leq 1 \text{ for arbitrary stopping times } \tau.$$

For more details on inference at arbitrary stopping times, we refer the reader to Howard et al. [2020, 2021], Grünwald et al. [2019], Ramdas et al. [2020].

6. A brief selective history on betting and its mathematical applications

From a purely statistical perspective, this paper could be viewed as tackling the problem of deriving sharp confidence sets for means of bounded random variables. In this pursuit, we find that a technique with excellent empirical performance happens to have strong connections to the topics of betting and gambling. While we provide a more detailed discussion in Section F, here we briefly summarize some of the ways in which betting ideas have appeared in and shaped probability, statistical inference, information theory, and online learning, in the broad context of our paper.

- **Probability**: The 1939 PhD thesis of Ville [1939] brought betting and martingales to the forefront of modern probability theory, by giving actionable interpretations to Kolmogorov's newly developed measure-theoretic probability, and dealing a near-fatal blow to the theory of collectives by von Mises. Ville showed that for *any* event A of probability measure zero (like sequences violating the law of large numbers), he could design an explicit betting strategy that never bets more than it has, whose wealth (a test martingale) grows without to infinity if the event A occurs. Ville worked with binary sequences, but his result holds more generally; see Shafer and Vovk [2001].

  One may view Ville's result as a theorem in measure-theoretic probability theory; what he effectively proved was: the event that a test (super)martingale exceeds $1/\alpha$ has probability at most α (Ville's inequality in this paper). This holds for any $\alpha \in [0,1]$, treating $1/0 \equiv +\infty$, with the $\alpha = 0$ case being the most remarkable part. But Ville's result is also an axiomatic building block for *game-theoretic probability* [Vovk, 1993, Shafer and Vovk, 2001, 2019]. Many classical results in probability can been derived in completely game-theoretic terms [Shafer and Vovk, 2001, 2019]. The capital processes used for deriving CSs are of the same form as those used to derive these foundational theorems of game-theoretic probability, despite the two goals being quite different.

- **Statistical inference**: The famous book of Wald [1945] was the first to thoroughly present and study sequential hypothesis testing. Despite not being presented in this way by Wald, we know in hindsight that the sequential probability ratio test (SPRT) is quite centrally based on the fact that the likelihood ratio is a nonnegative martingale. Two decades later, Robbins and colleagues built on Wald's sequential testing work in several ways, including to estimation via confidence sequences [Darling and Robbins, 1967a,b,c, Robbins and Siegmund, 1968, 1969, 1970, 1972, 1974, Robbins, 1970, Lai, 1976]. The recent work of Howard et al. [2020, 2021], Ramdas et al. [2021], Wasserman et al. [2020] extends the early work of Wald, Robbins and colleagues to a broader class of problems using exponential supermartingales and "e-processes", which can be seen as nonparametric, composite generalizations of the SPRT martingale. Connections between *betting* and the works of Wald, Robbins et al., and Howard et al. are implicit in those works, but can now be seen in hindsight, and our paper makes these connections explicit.

- **Information theory**: Working in the new field of information theory, Kelly Jr [1956] made direct connections to betting by showing that the capacity of a channel (itself fundamentally related to entropy and the Kullback-Leibler divergence) is given by the maximal rate of growth of wealth of a gambler in a simple game with iid Bernoulli(p) observations and known p. Breiman [1961] generalized Kelly's results significantly, and Krichevsky and Trofimov [1981] extended these results beyond the case of known p using a mixture method. Thomas Cover's interest in these techniques spans several decades [Cover, 1974, 1984, 1987, Bell and Cover, 1980, 1988], culminating in his famous universal portfolio algorithm [Cover, 1991]. The results of Krichevsky-Trofimov and Cover are essentially regret inequalities, leading directly to the final subfield below.

- **Online learning**: The techniques of Krichevsky, Trofimov and Cover found extensive applications to *sequential prediction with the logarithmic loss* [Cesa-Bianchi and Lugosi, 2006]. Here, one derives *regret inequalities* for the total loss accumulated when predicting the next observation from a potentially adversarial sequence. This problem is fundamentally connected to online convex optimization, for which Orabona and colleagues use parameter-free betting algorithms to derive regret inequalities [Orabona and Pal, 2016, Orabona and Tommasi, 2017, Jun et al., 2017, Cutkosky and Orabona, 2018, Jun and Orabona, 2019]. Rakhlin and Sridharan [2017] articulated a deep connection between martingale concentration and deterministic regret inequalities, and Jun and Orabona [2019, Section 7.1] derive concentration bounds for the general setting of Banach space-valued observations with sub-exponential noise.

7. Summary

Nonparametric confidence sequences are particularly useful in sequential estimation because they enable valid inference at arbitrary stopping times, but they are underappreciated as powerful tools to provide accurate inference even at fixed times. Recent work [Howard et al., 2020, 2021] has developed several time-uniform generalizations of the Cramer-Chernoff technique utilizing "line-crossing" inequalities and using various

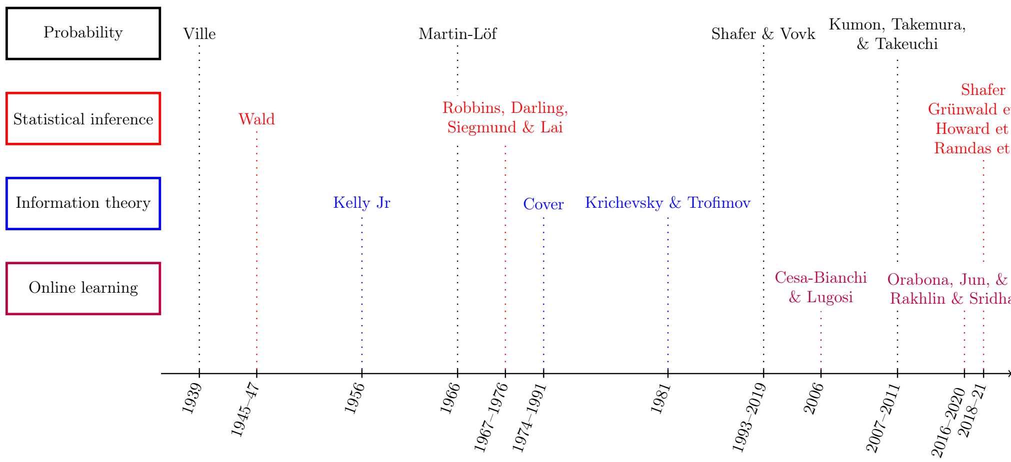

Figure 9. A brief selective history of betting ideas appearing (often implicitly) in various literatures. As discussed further in Section F, these subfields have rarely cited each other, but ideas are now beginning to permeate. Several authors did not explicitly use the language of betting, and their inclusion above is due to reinterpreting their work in hindsight.

variants of Robbins' method of mixtures (discrete mixtures, conjugate mixtures and stitching) to convert them to "curve-crossing" inequalities.

This work adds new techniques to the toolkit: to complement the aforementioned mixture methods, we develop a "predictable plug-in" approach. When coupled with existing nonparametric supermartingales, it yields (for example) computationally efficient empirical-Bernstein confidence sequences. One of our major contributions is to thoroughly develop the theory and methodology for a new nonnegative martingale approach to estimating means of bounded random variables in both with- and without-replacement settings. These convincingly outperform all existing published work that we are aware of, for CIs and CSs, both with and without replacement.

Our methods are particularly easy to interpret in terms of evolving capital processes and sequential testing by betting [Shafer, 2021] but we go much further by developing powerful and efficient betting strategies that lead to state-of-the-art variance-adaptive confidence sets that are significantly tighter than past work in all considered settings. In particular, Shafer espouses *complementary* benefits of such approaches, ranging from improved scientific communication, ties to historical advances in probability, and reproducibility via continued experimentation (also see Grünwald et al. [2019]), but our focus here has been on developing a new state of the art for a set of classical, fundamental problems.

There appear to be nontrivial connections to online learning theory [Kotłowski et al., 2010, Kumon et al., 2011, Orabona and Tommasi, 2017, Cutkosky and Orabona, 2018], and to empirical and dual likelihoods (see Section E.6 and an extended historical review of betting in Section F). The reductions from regret inequalities to concentration bounds described in Rakhlin and Sridharan [2017] and Jun and Orabona [2019] are fascinating, but existing published bounds are loose in the constants and not competitive in practice compared to our direct approach. Exploring deeper connections may yield other confidence sequences or betting strategies.

It is clear to us, and hopefully to the reader as well, that the ideas behind this work (adaptive statistical inference by betting) form the tip of the iceberg—they lead to powerful, efficient, nonasymptotic, nonparametric inference and can be adapted to a range of other problems. As just one example, let $\mathcal{P}^{p,q}$ represent the set of all continuous distributions such that the p-quantile of X_t, conditional on the past, is equal to q. This is also a nonparametric, convex set of distributions with no common reference measure. Nevertheless, for any predictable (λ_i), it is easy to check that

$$M_t = \prod_{i=1}^{t}(1 + \lambda_i(\mathbf{1}_{X_i \leq q} - p))$$

is a test martingale for $\mathcal{P}^{p,q}$. Setting $p = 1/2$ and $q = 0$, for example, we can sequentially test if the median of the underlying data distribution is the origin. The continuity assumption can be relaxed, and this test can be inverted to get a confidence sequence for any quantile. We do not pursue this idea further in the current paper because the recent (rather different) nonnegative martingale methods of Howard and Ramdas [2022] already provide a challenging benchmark for that problem. Typically, one test martingale-based method cannot uniformly dominate another, and the large gains in this paper were made possible because all previous published approaches implicitly or explicitly employed test *super*martingales, while we employ test martingales that are computationally simple to implement.

To conclude, we opine that "game-theoretic statistical inference" is in its nascency, and we expect much theoretical and practical progress in coming years. We hope the reader shares our excitement in this regard.

Data availability statement

No new data were generated or analysed in support of this research.

Acknowledgments. AR acknowledges funding from NSF DMS 1916320, an Adobe faculty research award and an NSF DMS (CAREER) 1945266. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC) [Nystrom et al., 2015]. The authors thank Harrie Hendriks, Philip Stark, Francesco Orabona, Kwang-Sung Jun, Nikos Karampatziakis and Arun Kuchibhotla for discussions on an early preprint, as well as Glenn Shafer, Vladimir Vovk and Peter Grünwald for broader discussions.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Theodore Anderson. Confidence limits for the expected value of an arbitrary bounded random variable with a continuous distribution function. Technical report, Stanford University Department of Statistics, 1969.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Tuning bandit algorithms in stochastic environments. In *Algorithmic Learning Theory*, 2007.

Akshay Balsubramani. Sharp finite-time iterated-logarithm martingale concentration. *arXiv:1405.2639*, 2014.

Rémi Bardenet and Odalric-Ambrym Maillard. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385, 2015.

Robert Bell and Thomas M Cover. Game-theoretic optimal portfolios. *Management Science*, 34(6):724–733, 1988.

Robert M Bell and Thomas M Cover. Competitive optimality of logarithmic investment. *Mathematics of Operations Research*, pages 161–166, 1980.

George Bennett. Probability Inequalities for the Sum of Independent Random Variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

Vidmantas Bentkus. On Hoeffding's inequalities. *The Annals of Probability*, 32(2): 1650–1673, 2004.

Vidmantas Bentkus, N Kalosha, and M Van Zuijlen. On domination of tail probabilities of (super) martingales: explicit bounds. *Lithuanian Mathematical Journal*, 46 (1):1–43, 2006.

Sergei Bernstein. *Theory of probability.* Gastehizdat Publishing House, 1927.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: a nonasymptotic theory of independence.* Oxford University Press, Oxford, 1st edition, 2013.

Leo Breiman. Optimal gambling systems for favorable games. *Berkeley Symposium on Mathematical Statistics and Probability*, 4.1(65-78), 1961.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, pages 1516–1541, 2013.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games.* Cambridge university press, 2006.

Thomas M Cover. Universal gambling schemes and the complexity measures of Kolmogorov and Chaitin. *Technical Report, no. 12*, 1974.

Thomas M Cover. An algorithm for maximizing expected log investment return. *IEEE Transactions on Information Theory*, 30(2):369–373, 1984.

Thomas M Cover. Log optimal portfolios. In *Chapter in "Gambling Research: Gambling and Risk Taking," Seventh International Conference*, volume 4, 1987.

Thomas M Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–29, 1991.

Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in Banach spaces. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, 2018.

D. A. Darling and Herbert Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, 58(1):66–68, 1967a.

D. A. Darling and Herbert Robbins. Inequalities for the Sequence of Sample Means. *Proceedings of the National Academy of Sciences*, 57(6):1577–1580, 1967b.

D. A. Darling and Herbert Robbins. Iterated Logarithm Inequalities. *Proceedings of the National Academy of Sciences*, 57(5):1188–1192, 1967c.

A Philip Dawid. Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–290, 1984.

A Philip Dawid. Prequential analysis. *Encyclopedia of Statistical Sciences*, 1:464–470, 1997.

Victor H. de la Peña, Michael J. Klass, and Tze Leung Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *The Annals of Probability*, 32(3):1902–1933, 2004. ISSN 0091-1798, 2168-894X.

Victor H. de la Peña, Michael J. Klass, and Tze Leung Lai. Pseudo-maximization and self-normalized processes. *Probability Surveys*, 4:172–192, 2007.

Victor H. de la Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: limit theory and statistical applications.* Springer, Berlin, 2009.

Joseph Leo Doob. *Stochastic Processes*, volume 10. New York Wiley, 1953.

Xiequan Fan, Ion Grama, and Quansheng Liu. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20(1):1–22, 2015.

Peter Grünwald. *The minimum description length principle.* MIT press, 2007.

Peter Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing. *arXiv preprint arXiv:1906.07801*, 2019.

Peter Hall and Barbara La Scala. Methodology and algorithms of empirical likelihood. *International Statistical Review*, pages 109–127, 1990.

Harrie Hendriks. Test martingales for bounded random variables. *arXiv preprint arXiv:1801.09418*, 2018.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *COLT*, pages 67–79. Citeseer, 2010.

Steven R. Howard and Aaditya Ramdas. Sequential estimation of quantiles with applications to A/B testing and best-arm identification. *Bernoulli*, 28(3):1704 – 1728, 2022.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

Kwang-Sung Jun and Francesco Orabona. Parameter-free online convex optimization with sub-exponential noise. In *Conference on Learning Theory*, pages 1802–1823. PMLR, 2019.

Kwang-Sung Jun, Francesco Orabona, Stephen Wright, and Rebecca Willett. Improved strongly adaptive online learning using coin betting. In *Artificial Intelligence and Statistics*, pages 943–951. PMLR, 2017.

Emilie Kaufmann and Wouter M. Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021.

Michael Kearns and Lawrence Saul. Large deviation methods for approximate probabilistic inference. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, pages 311–319, 1998.

John L Kelly Jr. A new interpretation of information rate. *Bell System Technical Journal*, 35(4):917–926, 1956.

Wojciech Kotłowski, Peter Grünwald, and Steven De Rooij. Following the flattened leader. In *Conference on Learning Theory*, pages 106–118. Citeseer, 2010.

Raphail Krichevsky and Victor Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207, 1981.

Arun K Kuchibhotla and Qinqing Zheng. Near-optimal confidence sequences for bounded random variables. *International Conference on Machine Learning*, 2021.

Masayuki Kumon, Akimichi Takemura, and Kei Takeuchi. Sequential optimizing strategy in multi-dimensional bounded forecasting games. *Stochastic Processes and their Applications*, 121(1):155–183, 2011.

Tze Leung Lai. On Confidence Sequences. *The Annals of Statistics*, 4(2):265–280, March 1976. ISSN 0090-5364, 2168-8966.

Erik Learned-Miller and Philip S Thomas. A new confidence interval for the mean of a bounded random variable. *arXiv preprint arXiv:1905.06208*, 2019.

Qiang Jonathan Li. *Estimation of mixture models*. PhD thesis, Yale University, 1999.

Gary Lorden and Moshe Pollak. Nonanticipating estimation applied to sequential analysis and changepoint detection. *The Annals of Statistics*, 33(3):1422–1454, June 2005. ISSN 0090-5364, 2168-8966.

Per Martin-Löf. The definition of random sequences. *Information and control*, 9(6): 602–619, 1966.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the Conference on Learning Theory*, 2009.

Michael McKerns, Leif Strand, Tim Sullivan, Alta Fang, and Michael Aivazis. Building a framework for predictive science. *Proceedings of the 10th Python in Science Conference*, 2011.

Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical Bernstein stopping. In *Proceedings of the 25th International Conference on Machine Learning*, pages 672–679. ACM, 2008.

Per Aslak Mykland. Dual likelihood. *The Annals of Statistics*, pages 396–421, 1995.

Nicholas A. Nystrom, Michael J. Levine, Ralph Z. Roskies, and J. Ray Scott. Bridges: A uniquely flexible hpc resource for new communities and data analytics. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, XSEDE '15, pages 30:1–30:8, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3720-5. doi: 10.1145/2792745.2792775. URL http://doi.acm.org/10.1145/2792745.2792775.

Francesco Orabona and David Pal. Coin betting and parameter-free online learning. *Advances in Neural Information Processing Systems*, 29:577–585, 2016.

Francesco Orabona and Tatiana Tommasi. Training deep networks without learning rates through coin betting. In *Advances in Neural Information Processing Systems*, pages 2160–2170, 2017.

Art B Owen. *Empirical likelihood*. CRC press, 2001.

My Phan, Philip S Thomas, and Erik Learned-Miller. Towards practical mean bounds for small samples. *International Conference on Machine Learning*, 2021.

Alexander Rakhlin and Karthik Sridharan. On equivalence of martingale tail bounds and deterministic regret inequalities. In *Conference on Learning Theory*, pages 1704–1722. PMLR, 2017.

Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*, 2020.

Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter M Koolen. Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, 2021.

Jorma Rissanen. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, 30(4):629–636, 1984.

Jorma Rissanen. *Stochastic Complexity in Statistical Inquiry*, volume 15. World Scientific, 1998.

Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.

Herbert Robbins and David Siegmund. Iterated logarithm inequalities and related statistical procedures. In *Mathematics of the Decision Sciences, Part II*, pages 267–279. American Mathematical Society, Providence, 1968.

Herbert Robbins and David Siegmund. Probability distributions related to the law of the iterated logarithm. *Proc. of the National Academy of Sciences*, 62(1):11–13, January 1969. ISSN 0027-8424.

Herbert Robbins and David Siegmund. Boundary crossing probabilities for the Wiener process and sample sums. *The Annals of Mathematical Statistics*, 41(5): 1410–1429, 1970.

Herbert Robbins and David Siegmund. A class of stopping rules for testing parametric hypotheses. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, volume 4, pages 37–41, 1972.

Herbert Robbins and David Siegmund. The expected sample size of some tests of power one. *The Annals of Statistics*, 2(3):415–436, May 1974.

Robert J Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, pages 39–48, 1974.

Glenn Shafer. The language of betting as a strategy for statistical and scientific communication. *Journal of the Royal Statistical Society, Series A*, 2021.

Glenn Shafer and Vladimir Vovk. *Probability and Finance: It's Only a Game!* John Wiley & Sons, February 2001. ISBN 978-0-471-46171-5.

Glenn Shafer and Vladimir Vovk. *Game-Theoretic Foundations for Probability and Finance*. John Wiley & Sons, 2019.

Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test Martingales, Bayes Factors and p-Values. *Statistical Science*, 26(1):84–101, 2011.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Philip B Stark. Sets of half-average nulls generate risk-limiting audits: SHANGRLA. In *International Conference on Financial Cryptography and Data Security*, pages 319–336. Springer, 2020.

Jean Ville. *Étude Critique de la Notion de Collectif.* PhD thesis, Paris, 1939.

Vladimir Vovk. Testing randomness online. *Statistical Science*, 36(4):595–611, 2021.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Vladimir G Vovk. A logic of probability, with application to the foundations of statistics. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55 (2):317–341, 1993.

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

Abraham Wald. *Sequential Analysis.* John Wiley & Sons, New York, 1947.

Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 2020. ISSN 0027-8424.

Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. *Advances in Neural Information Processing Systems*, 33, 2020.

Shengjia Zhao, Enze Zhou, Ashish Sabharwal, and Stefano Ermon. Adaptive concentration inequalities for sequential decision problems. In *30th Conference on Neural Information Processing Systems*, 2016.