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

Abstract. We derive confidence intervals (CIs) and confidence sequences (CSs) for the classical problem of estimating a bounded mean. Our approach generalizes and improves on the celebrated Chernoff method, yielding the best closed-form "empirical-Bernstein" CSs and CIs (converging exactly to the oracle Bernstein width) as well as non-closed-form "betting" CSs and CIs. Our method combines new composite nonnegative (super)martingales with Ville's maximal inequality, with strong connections to testing by betting and the method of mixtures. We also show how these ideas can be extended to sampling without replacement. In all cases, our bounds are adaptive to the unknown variance, and empirically vastly outperform prior approaches, establishing 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

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

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.

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

‡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.

paper by Shafer (2021), but from a purely mathematical viewpoint, our results are extensions of a unified supermartingale approach towards nonparametric concentration 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.

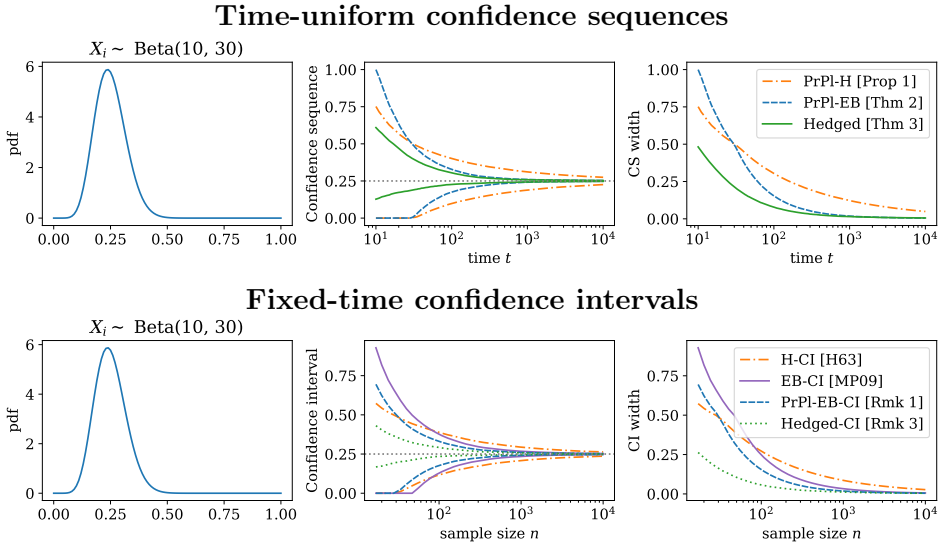


Figure 1. Time-uniform 95% confidence sequences (upper row) and fixed-time 95% confidence intervals (lower row) for the mean of independent and idenically distributed (iid) draws from a Beta(10, 30) distribution (unknown to the methods). The betting approaches (Hedged and Hedged-CI) adapt to both the 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 one is 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

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

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 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.

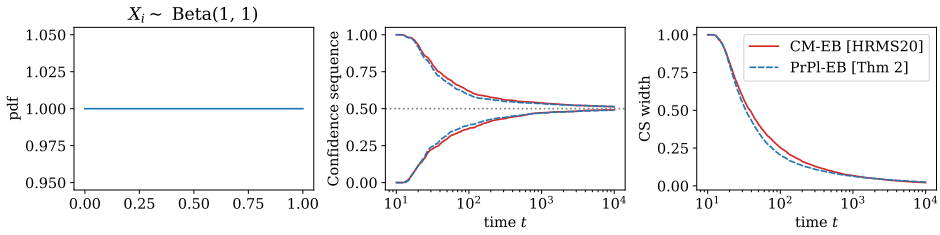


Figure 2. Empirical Bernstein CSs produced via a predictable plug-in (PrPl) with $(\lambda_t)_{t=1}^{\infty}$ from (15) match (or slightly improve) those obtained via conjugate mixtures (CM) by Howard et al. (2021); the former is closed-form, but the latter is not and requires numerical methods.

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 $X_1, \ldots, X_n$ 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 $(\widetilde{X}_{1,b}, \ldots, \widetilde{X}_{n,b})$ and correspondingly compute $\widetilde{M}_{n,b}^{\text{PrPl-EB}}(m)$, one for each permutation $b \in \{1, \ldots, B\}$. 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).

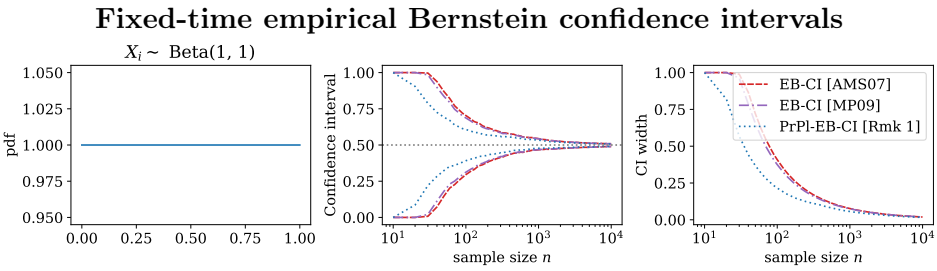


Figure 3. Our predictable plug-in (PrPl) empirical Bernstein (EB) CI is significantly tighter than those of Maurer and Pontil (2009) and Audibert et al. (2007).

Table 1. Below, we think of $\log x$ as $\log(x+1)$ to avoid trivialities. The claimed rates are easily checked. 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 which is why we treat $\log\log t$ as a constant when expressing the rate for W_t. The iterated logarithm pa

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

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 $X_1, X_2, \ldots$ 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* $\mathcal{F}_{t-1} = \sigma(X_1, \ldots, X_{t-1})$.

(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 $X_1, \ldots, X_{t-1}$. The adversary picks a distribution $Q_t \in \mathcal{Q}^\mu$, which could depend on $X_1, \ldots, X_{t-1}$ 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 $(0, 1/g, 2/g, \ldots, (g-1)/g, 1)$ 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

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

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

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

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.

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.*

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.*

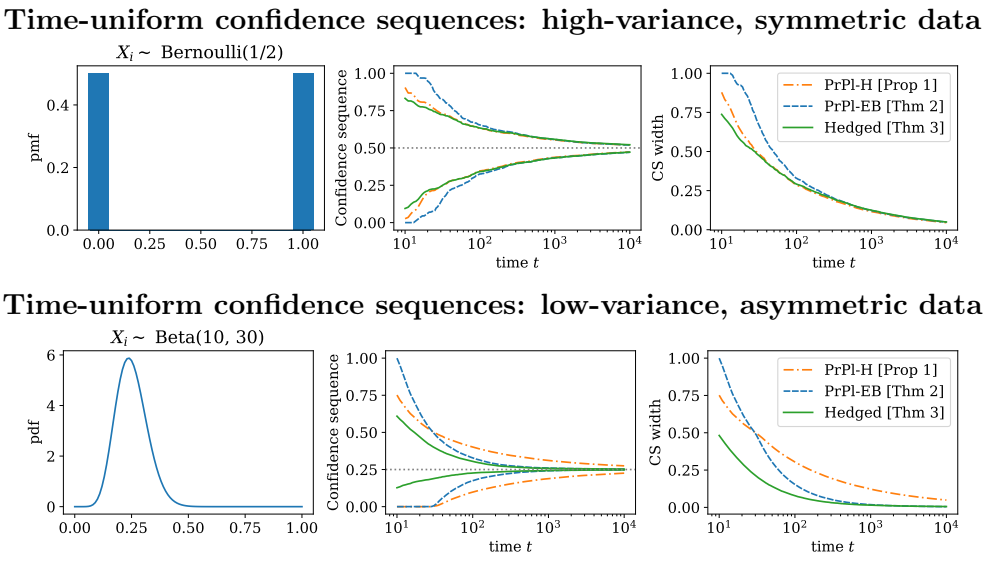


Figure 4. Predictable plug-in Hoeffding, empirical Bernstein, and hedged capital CSs under two distributional scenarios. Notice that the latter roughly matches the others in the Bernoulli$(1/2)$ case, but shines in the low-variance, asymmetric scenario.

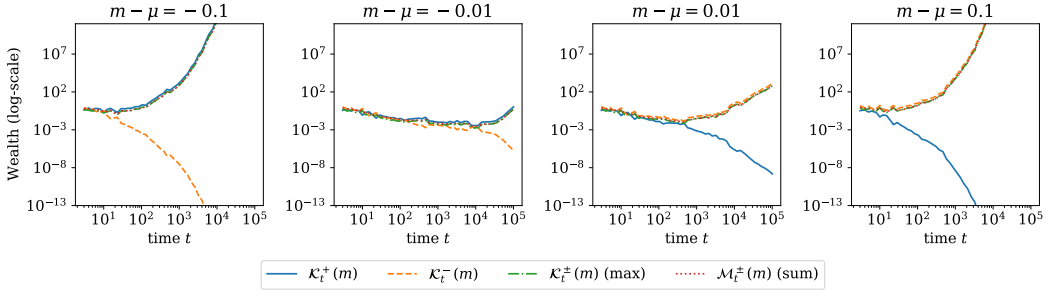


Figure 5. A comparison of capital processes $\mathcal{K}_t^+(m)$, $\mathcal{K}_t^-(m)$, the hedged capital process $\mathcal{K}_t^\pm(m)$, and its upper-bounding nonnegative martingale, $\mathcal{M}_t^\pm(m)$ under four alternatives (from left to right): $m \ll \mu$, $m < \mu$, $m > \mu$, $m \gg \mu$. 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 $m < \mu$.

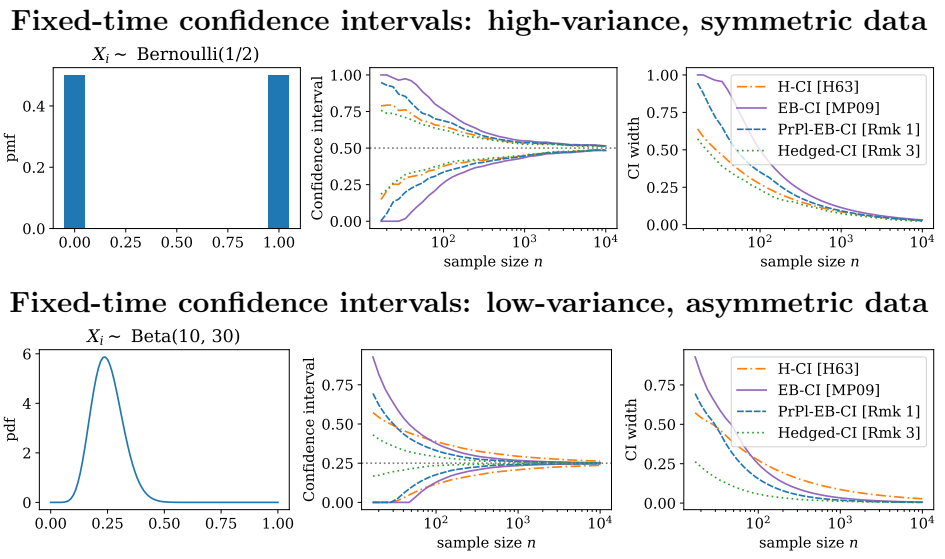


Figure 6. Hoeffding (H), empirical Bernstein (EB), and hedged capital CIs under two distributional scenarios. Similar to the time-uniform setting, the betting approach tends to outperform the other bounds, especially for low-variance, asymmetric data.

Similar to the discussion after Remark 1, if $X_1, \ldots, X_n$ 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),
- 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

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

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 $(\mu_t^{\text{WoR}})_{t\in[N]}$ where

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

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

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

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

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

with $\mathcal{K}_0^{\text{WoR}}(m) := 1$. 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:

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

for some predictable $(\lambda_t^+(m))_{t=1}^N$ and $(\lambda_t^-(m))_{t=1}^N$ taking values in $[0, 1/m_t^{\mathrm{WoR}}]$ and $[0, 1/(1-m_t^{\mathrm{WoR}})]$ at time t, respectively. Using $\left(\mathcal{K}_t^{\pm,\mathrm{WoR}}(m)\right)_{t=0}^{\infty}$ 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 $X_1, X_2, \ldots X_N$ 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*

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

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

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

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

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

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

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

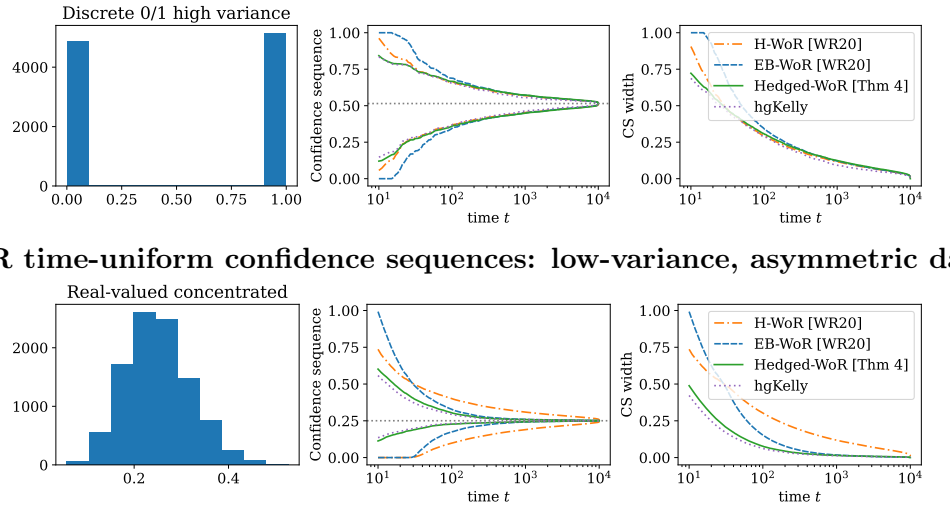


Figure 7. Without-replacement betting CSs versus the predictable plug-in supermartingale-based CSs (Waudby-Smith and Ramdas, 2020). Similar to the with-replacement case, the betting approach matches or vastly outperforms past state-of-the art methods.

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 $\bigcap_{i \leq n} \mathfrak{B}_i^{\pm,\mathrm{WoR}}$, 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.*

REMARK 5. *For some values of m near 0 or 1, m_t^{WoR} could lie outside of $[0,1]$, leading $\mathcal{K}_t^{\pm,\mathrm{WoR}}(m)$ to potentially be negative. However, it is impossible for $\mathcal{K}_t^{\pm,\mathrm{WoR}}(\mu)$ to be negative since $\mu_t \in [0,1]$ always. In fact, a negative m_t implies that the value of m being tested is impossible, and thus one can reject that null immediately. In particular, when running our method, one can instead use the modified capital process*

$$\widetilde{\mathcal{K}}_t^{\pm,\mathrm{WoR}}(m) := |\mathcal{K}_t^{\pm,\mathrm{WoR}}(m)|/\mathbf{1}(m_t \in [0,1]) \tag{35}$$

which takes on the value $+\infty$ if the denominator evaluates to zero. Note that $\widetilde{\mathcal{K}}_t^{\pm,\mathrm{WoR}}(\mu)$ still forms a nonnegative martingale since its denominator is always one when $m = \mu$.

Notice that constructing a WoR test martingale only relies on changing the fixed conditional mean μ to the time-varying conditional mean $\mu_t^{\mathrm{WoR}} := \frac{N\mu - \sum_{i=1}^{t-1} X_i}{N-t+1}$ and now designing $(-1/(1-\mu_t^{\mathrm{WoR}}), 1/\mu_t^{\mathrm{WoR}})$-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.

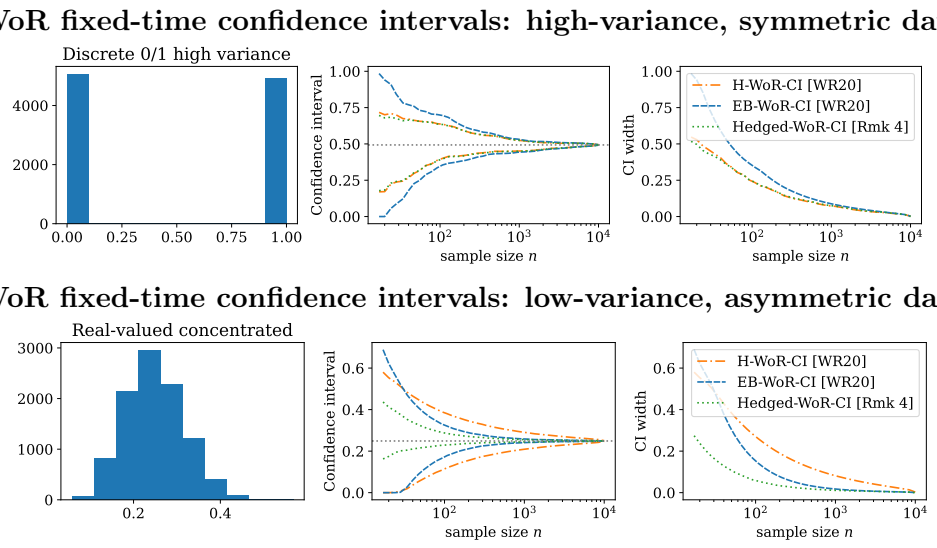


Figure 8. WoR analogue of the hedged capital CI versus the WoR Hoeffding- and empirical Bernstein-type CIs (Waudby-Smith and Ramdas, 2020). Similar to with-replacement, the betting approach has excellent performance.

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 $\mathcal{K}_t(m)$,

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

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

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

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

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

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 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 (1947) 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.

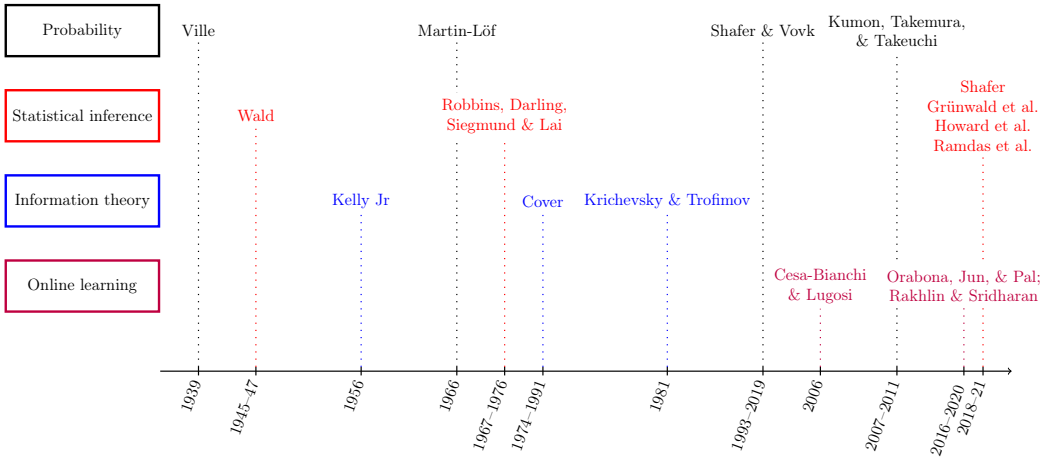


Figure 9. A brief selective history of betting ideas appearing (often implicitly) in various literatures. As discussed further in Section F, these subfields have rarely cited each other, but ideas are now beginning to permeate. Several authors did not explicitly use the language of betting, and their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

7. Summary

Nonparametric confidence sequences are particularly useful in sequential estimation because they enable valid inference at arbitrary stopping times, but they are un-

derappreciated 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 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

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

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.

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 J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL `https://proceedings.neurips.cc/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf`.

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.

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.

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.

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.

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

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

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.

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.

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.

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.

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.

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

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.

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.

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, 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.

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.

A. Proofs of main results

We first introduce a lemma which will aid in the proofs to follow.

LEMMA 1 (PREDICTABLE PLUG-IN CHERNOFF SUPERMARTINGALES). *Suppose that $X_1, X_2, \cdots \sim P$, and for some μ, v_t and $\psi(\lambda)$, we have that for any $\lambda \in \Lambda \subseteq \mathbb{R}$,*

$$\mathbb{E}_P\left[\exp(\lambda(X_t - \mu) - v_t\psi(\lambda)) \mid \mathcal{F}_{t-1}\right] \leq 1 \quad \textit{for each } t \geq 1\ . \tag{36}$$

Then, for any Λ-valued sequence $(\lambda_t)_{t=1}^{\infty}$ that is predictable with respect to $\mathcal{F}$,

$$M_t^{\psi}(\mu) := \prod_{i=1}^{t} \exp\left(\lambda_i(X_i - \mu) - v_i\psi(\lambda_i)\right)$$

forms a test supermartingale with respect to $\mathcal{F}$.

PROOF. Writing out the conditional expectation of M_t^{ψ} for any $t \geq 2$,

$$\begin{aligned}
\mathbb{E}\left(M_t^{\psi}(\mu) \mid \mathcal{F}_{t-1}\right) &= \mathbb{E}\left(\prod_{i=1}^{t} \exp\left(\lambda_i(X_i - \mu) - v_i\psi(\lambda_i)\right) \,\middle|\, \mathcal{F}_{t-1}\right) \\
&\overset{(i)}{=} \prod_{i=1}^{t-1} \exp\left(\lambda_i(X_i - \mu) - v_i\psi(\lambda_i)\right) \underbrace{\mathbb{E}\left[\exp\left(\lambda_t(X_t - \mu) - v_t\psi(\lambda_t)\right) \mid \mathcal{F}_{t-1}\right]}_{\leq 1 \text{ by assumption}} \\
&= M_{t-1}^{\psi}(\mu),
\end{aligned}$$

where (i) follows from the fact that $\exp\left(\lambda_i(X_i - \mu) - v_i\psi(\lambda_i)\right)$ is $\mathcal{F}_{t-1}$-measurable for $i \leq t-1$. Since $\mathcal{F}_0$ was assumed to be trivial, for M_1 we have that

$$\mathbb{E}[M_1^{\psi}(\mu)|\mathcal{F}_0] = \underbrace{\mathbb{E}\left[\exp\left(\lambda_1(X_1 - \mu) - v_1\psi(\lambda_1)\right)\right]}_{\leq 1 \text{ by assumption}},$$

which completes the proof. □

A.1. Proof of Proposition 1

The proof proceeds in three steps. First, apply a standard MGF bound by Hoeffding (1963). Second, we apply Lemma 1. Finally, we apply Theorem 1 to obtain a CS and take a union bound.

Step 1. By Hoeffding (1963), we have that $\mathbb{E}\left[\exp(\lambda_t(X_t - \mu) - \psi_H(\lambda_t)) \mid \mathcal{F}_{t-1}\right] \leq 1$ since $X_t \in [0,1]$ almost surely and since λ_t is $\mathcal{F}_{t-1}$-measurable.

Step 2. By Step 1 and Lemma 1, we have that

$$M_t^{\text{PrPl-H}}(\mu) := \prod_{i=1}^{t} \exp\left(\lambda_i(X_i - \mu) - \psi_H(\lambda_i)\right)$$

forms a test supermartingale.

Step 3. By Step 2 combined with Theorem 1, we have that

$$\begin{aligned} &P\left(\exists t \geq 1 : \mu \leq \frac{\sum_{i=1}^t \lambda_i X_i}{\sum_{i=1}^t \lambda_i} - \frac{\log(1/\alpha) + \sum_{i=1}^t \psi_H(\lambda_i)}{\sum_{i=1}^t \lambda_i}\right) \\ &= P\left(\exists t \geq 1 : M_t^{\text{PrPl-H}}(\mu) \geq 1/\alpha\right) \leq \alpha. \end{aligned}$$

Applying the same bound to $(-X_t)_{t=1}^\infty$ with mean $-\mu$ and taking a union bound, we have the desired result,

$$P\left(\exists t \geq 1 : \mu \notin \left(\frac{\sum_{i=1}^t \lambda_i X_i}{\sum_{i=1}^t \lambda_i} \pm \frac{\log(2/\alpha) + \sum_{i=1}^t \psi_H(\lambda_i)}{\sum_{i=1}^t \lambda_i}\right)\right) \leq \alpha,$$

which completes the proof. □

A.2. Proof of Theorem 2

By Lemma 1 combined with Theorem 1, it suffices to prove that

$$\mathbb{E}_P\left[\exp\left\{\lambda_t(X_t - \mu) - v_t\psi_E(\lambda_t)\right\} \mid \mathcal{F}_{t-1}\right] \leq 1.$$

For succinctness, denote

$$Y_t := X_t - \mu \quad \text{and} \quad \delta_t := \widehat{\mu}_t - \mu.$$

Note that $\mathbb{E}_P(Y_t \mid \mathcal{F}_{t-1}) = 0$. It then suffices to prove that for any $[0,1)$-bounded, $\mathcal{F}_{t-1}$- measurable $\lambda_t \equiv \lambda_t(X_1^{t-1})$,

$$\mathbb{E}\left[\exp\left\{\lambda_t Y_t - 4(Y_t - \delta_{t-1})^2\psi_E(\lambda_t)\right\} \,\middle|\, \mathcal{F}_{t-1}\right] \leq 1.$$

Indeed, in the proof of Proposition 4.1 in Fan et al. (2015), $\exp\{\xi\lambda - 4\xi^2\psi_E(\lambda)\} \leq 1 + \xi\lambda$ for any $\lambda \in [0,1)$ and $\xi \geq -1$. Setting $\xi := Y_t - \delta_{t-1} = X_t - \widehat{\mu}_{t-1}$,

$$\begin{aligned} &\mathbb{E}\left[\exp\left\{\lambda_t Y_t - 4(Y_t - \delta_{t-1})^2\psi_E(\lambda_t)\right\} \,\middle|\, \mathcal{F}_{t-1}\right] \\ &= \mathbb{E}\Big[\exp\Big\{\lambda_t(Y_t - \delta_{t-1}) - 4(Y_t - \delta_{t-1})^2\psi_E(\lambda_t)\Big\} \,\Big|\, \mathcal{F}_{t-1}\Big]\exp(\lambda_t\delta_{t-1}) \\ &\leq \mathbb{E}\Big[1 + (Y_t - \delta_{t-1})\lambda_t \mid \mathcal{F}_{t-1}\Big]\exp(\lambda_t\delta_{t-1}) \overset{(i)}{=} \mathbb{E}\big[1 - \delta_{t-1}\lambda_t \mid \mathcal{F}_{t-1}\big]\exp(\lambda_t\delta_{t-1}) \overset{(ii)}{\leq} 1, \end{aligned}$$

where equality (i) follows from the fact that Y_t is conditionally mean zero, and inequality (ii) follows from the inequality $1 - x \leq \exp(-x)$ for all $x \in \mathbb{R}$. This completes the proof. □

A.3. Proof of Proposition 2

We proceed by proving $(d) \implies (c) \implies (b) \implies (a) \implies (d)$.

Proof of $(d) \implies (c)$**.** This claim follows from the fact that for $\lambda \in (-1/(1-\mu), 1/\mu)$, we have that $(\lambda, \lambda, \dots)$ is a $(-1/(1-\mu), 1/\mu)$-valued predictable sequence.

Proof of $(c) \implies (b)$**.** By the assumption of (c), we have that for $\lambda = 0.5$, $\mathcal{K}_t(\mu)$ forms a test martingale. Furthermore, since $X_i, \mu \in [0,1]$ for each $i \in \{1, 2, \dots\}$, we have that $1 + 0.5(X_i - \mu) > 0$ almost surely for each i. Therefore, $(\mathcal{K}_t(\mu))_{t=1}^{\infty}$ is a strictly positive test martingale.

Proof of $(b) \implies (a)$**.** Suppose that there exists $\lambda \in \mathbb{R} \setminus \{0\}$ such that $\mathcal{K}_t(\mu)$ forms a strictly positive martingale. Then we must have

$$\begin{aligned}\mathcal{K}_{t-1}(\mu) &= \mathbb{E}\left(\mathcal{K}_t(\mu) \mid \mathcal{F}_{t-1}\right) \\ &= \mathcal{K}_{t-1}(\mu) \cdot \mathbb{E}\left(1 + \lambda(X_i - \mu) \mid \mathcal{F}_{t-1}\right) \\ &= \mathcal{K}_{t-1}(\mu) \cdot \left[1 + \lambda(\mathbb{E}(X_t \mid \mathcal{F}_{t-1}) - \mu)\right].\end{aligned}$$

Now since $\mathcal{K}_{t-1}(\mu) > 0$, we have that

$$1 + \lambda(\mathbb{E}(X_t \mid \mathcal{F}_{t-1}) - \mu) = 1.$$

Since $\lambda \neq 0$ by assumption, we have that $\mathbb{E}(X_t \mid \mathcal{F}_{t-1}) = \mu$ as required.

Proof of $(a) \implies (d)$**.** Let $(\lambda_t(\mu))_{t=1}^{\infty}$ be a $(-1/(1-\mu), 1/\mu)$-valued predictable sequence. Then $\mathcal{K}_t(\mu)$ is clearly nonnegative and $\mathcal{K}_0(\mu) = 1$ by definition. Writing out the conditional mean of the capital process for any $t \geq 1$,

$$\begin{aligned}\mathbb{E}\left(\mathcal{K}_t(\mu) \mid \mathcal{F}_{t-1}\right) &= \mathcal{K}_{t-1}(\mu) \cdot \mathbb{E}\left(1 + \lambda_t(\mu)(X_i - m) \mid \mathcal{F}_{t-1}\right) \\ &= \mathcal{K}_{t-1}(\mu) \cdot \left[1 + \lambda_t(\mu)(\mathbb{E}(X_i \mid \mathcal{F}_{t-1}) - \mu)\right] \\ &= \mathcal{K}_{t-1}(\mu),\end{aligned}$$

and thus $\mathcal{K}_t(\mu)$ forms a test martingale.

The proof of the final part of the proposition is simple. Let (M_t) be a test martingale for $\mathcal{P}^{\mu}$. Define $Y_t := M_t / M_{t-1}$ if $M_{t-1} > 0$, and as $Y_t := 0$ otherwise. Now note that $M_t = \prod_{i=1}^{t} Y_t$ and $\mathbb{E}_P[Y_t | \mathcal{F}_{t-1}] = 1$ for any $P \in \mathcal{P}^{\mu}$. In other words, every test martingale is a product of nonnegative random variables with conditional mean one. Now rewrite Y_t as $(1 + f_t(X_t))$ for some predictable function f_t. Since Y_t is nonnegative, we must have $f_t(X_t) \geq -1$, and since Y_t is conditional mean one, we must have $f_t(X_t)$ is conditional mean zero. Such a representation in fact holds true for any test martingale, and we have not yet used the fact that we are working with test martingales for $\mathcal{P}^{\mu}$. Now, the proof ends by noting that the only predictable functions f_t with the latter property under every $P \in \mathcal{P}^{\mu}$ has the form $\lambda_t(X_t - \mu)$ for some predictable λ_t; any nonlinear function of X_t would not have mean zero under *every distribution* with mean μ.

This completes the proof of Proposition 2 altogether. □

A.4. Proof of Proposition 3

We only prove the martingale part of the proposition, since the supermartingale aspect follows analogously, and as mentioned early in the paper, inequalities and equalities are meant in an almost sure sense.

First, it is easy to check that if (M_t) is a test martingale for $\mathcal{S}$, then M_t is the product of nonnegative conditionally unit mean terms, that is $M_t = \prod_{i=1}^t Y_i$ such that for all $S \in \mathcal{S}$, we have $\mathbb{E}_S[Y_i|\mathcal{F}_{i-1}] = 1$ and $Y_i \geq 0$. (Indeed, one can identify $Y_i := \frac{M_i}{M_{i-1}}\mathbf{1}_{M_{i-1}>0}$.) Now, define $Z_i' := Y_i - 1$, and note that $Z_i' \geq -1$, and $\mathbb{E}_S[Z_i'|\mathcal{F}_{t-1}] = 0$. Thus, M_t has been represented as $\prod_{i=1}^t (1 + Z_i')$. Now, the proof is completed by noting that any such Z_i' can be written as $\lambda_i Z_i$ for a predictable λ_i (this step is purely cosmetic). □

A.5. Proof of Theorem 3

First, we present Lemma 2 which establishes that the hedged capital process is a quasiconvex function of m (and thus has convex sublevel sets). We then invoke this lemma to prove the main result.

LEMMA 2. *Let* $\theta \in [0,1]$ *and*

$$\begin{aligned}\mathcal{K}_t^{\pm}(m) &:= \max\left\{\theta\mathcal{K}_t^{+}(m), (1-\theta)\mathcal{K}_t^{-}(m)\right\} \\ &\equiv \max\left\{\theta\prod_{i=1}^{t}(1+\lambda_i^{+}(m)\cdot(X_i-m)), (1-\theta)\prod_{i=1}^{t}(1-\lambda_i^{-}(m)\cdot(X_i-m))\right\}\end{aligned}$$

be the hedged capital process as in Section 4. Consider the $(1-\alpha)$ *confidence set of the same theorem,*

$$\mathfrak{B}_t^{\pm} \equiv \mathfrak{B}^{\pm}(X_1,\ldots,X_t) := \left\{m \in [0,1] : \mathcal{K}_t^{\pm}(m) < \frac{1}{\alpha}\right\}.$$

Then $\mathfrak{B}_t^{\pm}$ *is an interval on* $[0,1]$.

PROOF. Since sublevel sets of quasiconvex functions are convex, it suffices to prove that $\mathcal{K}_t^{\pm}(m)$ is a quasiconvex function of $m \in [0,1]$. The crux of the argument is: the product of nonnegative nonincreasing functions is quasiconvex, the product of nonnegative nondecreasing functions is also quasiconvex, and the maximum of quasiconvex functions is quasiconvex.

To elaborate, we will proceed in two steps. First, we use an induction argument to show that $\mathcal{K}_t^{+}(m)$ and $\mathcal{K}_t^{-}(m)$ are nonincreasing and nondecreasing, respectively, and hence quasiconvex. Finally, we note that $\mathcal{K}_t^{\pm}(m) := \max\left\{\theta\mathcal{K}_t^{+}(m), (1-\theta)\mathcal{K}_t^{-}(m)\right\}$ is a maximum of quasiconvex functions and is thus itself quasiconvex.

Step 1. First, since $\dot{\lambda}_t^{+}$ does not depend on m, we have that

$$1 + \lambda_t^{+}(m)(X_t - m) := 1 + \left(|\dot{\lambda}_t^{+}| \wedge \frac{c}{m}\right)(X_t - m)$$

is nonnegative and nonincreasing in m for each $t \in \{1, 2, \dots\}$. (To see this, consider the terms with and without truncation separately.) Suppose for the sake of induction that

$$\prod_{i=1}^{t-1} \big(1 + \lambda_i^+(m)(X_i - m)\big)$$

is nonnegative and nonincreasing in m. Then,

$$\begin{aligned}\mathcal{K}_t^+(m) &:= \prod_{i=1}^{t} \big(1 + \lambda_i^+(m)(X_i - m)\big) \\ &= \big(1 + \lambda_t^+(m)(X_t - m)\big) \cdot \prod_{i=1}^{t-1} \big(1 + \lambda_i^+(m)(X_i - m)\big)\end{aligned}$$

is a product of nonnegative and nonincreasing functions, and is thus itself nonnegative and nonincreasing. By a similar argument, $\mathcal{K}_t^-(m)$ is nonnegative and *nondecreasing*. $\mathcal{K}_t^+(m)$ and $\mathcal{K}_t^-(m)$ are thus both quasiconvex.

Step 2. Since the maximum of quasiconvex functions is quasiconvex, we infer that

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

is quasiconvex. In particular, the sublevel sets of quasiconvex functions is convex, and thus

$$\mathfrak{B}_t^\pm := \left\{m \in [0,1] : \mathcal{K}_t^\pm(m) < \frac{1}{\alpha}\right\}$$

is an interval, which completes the proof of Lemma 2. □

Proof (Theorem 3). The proof proceeds in three steps. First we show that $\mathcal{K}_t^\pm(\mu)$ is upper-bounded by test martingale. Second, we apply the 4-step procedure in Theorem 1 to get a CS for μ. Third and finally, we invoke Lemma 2 to conclude that the CS is indeed convex at each time t.

Step 1. We first upper bound $\mathcal{K}_t^\pm(m)$ as follows:

$$\begin{aligned}\mathcal{K}_t^\pm(m) &:= \max\big\{\theta\mathcal{K}_t^+(m), (1-\theta)\mathcal{K}_t^-(m)\big\} \\ &\le \theta\mathcal{K}_t^+(m) + (1-\theta)\mathcal{K}_t^-(m) =: \mathcal{M}_t^\pm(m).\end{aligned}$$

By Proposition 2, we have that $\mathcal{K}_t^+(\mu)$ and $\mathcal{K}_t^-(\mu)$ are test martingales for $\mathcal{P}$. For each $P \in \mathcal{P}$, writing out the conditional expectation of $\mathcal{M}_t^\pm(\mu)$ for any $t \ge 1$,

$$\begin{aligned}\mathbb{E}_P\left[\mathcal{M}_t^\pm(\mu) \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}_P\left[\theta\mathcal{K}_t^+(\mu) + (1-\theta)\mathcal{K}_t^-(\mu) \;\middle|\; \mathcal{F}_{t-1}\right] \\ &= \theta\mathbb{E}_P(\mathcal{K}_t^+(\mu) \mid \mathcal{F}_{t-1}) + (1-\theta)\mathbb{E}_P(\mathcal{K}_t^-(\mu) \mid \mathcal{F}_{t-1}) \\ &= \theta\mathcal{K}_{t-1}^+(\mu) + (1-\theta)\mathcal{K}_{t-1}^-(\mu) \\ &= \mathcal{M}_{t-1}^\pm(\mu),\end{aligned}$$

and $\mathcal{M}_0^\pm(\mu) = \theta\mathcal{K}_0^+(\mu) + (1-\theta)\mathcal{K}_0^-(\mu) = 1$. Therefore, $(\mathcal{M}_t^\pm(\mu))_{t=0}^\infty$ is a test martingale for $\mathcal{P}$.

Step 2. By Step 1 combined with Theorem 1 we have that

$$\mathfrak{B}_t^{\pm} := \left\{ m \in [0,1] : \mathcal{K}_t^{\pm}(m) < \frac{1}{\alpha} \right\}$$

forms a $(1-\alpha)$-CS for μ.

Step 3. Finally, by Lemma 2, we have that $\mathfrak{B}_t^{\pm}$ is an interval for each $t \in \{1, 2, \dots\}$, which completes the proof of Theorem 3. □

A.6. Proof of Lemma 3

Following the proof of Lemma 4.1 in Fan et al. (2015), we have that the function

$$f(x) := \begin{cases} \dfrac{\log(1+x) - x}{x^2/2} & x \in (-1, \infty) \setminus \{0\} \\ -1 & x = 0 \end{cases} \tag{37}$$

is an increasing and continuous function in x (note that $f(0)$ is defined as -1 because it is a removable singularity). For any $y \geq -m$ and $\lambda \in [0, 1/m)$ we have

$$\lambda y \geq -m\lambda > -1. \tag{38}$$

Combining (37) and (38), we have

$$\frac{\log(1+\lambda y) - \lambda y}{\lambda^2 y^2/2} \geq \frac{\log(1-m\lambda) + m\lambda}{\lambda^2 m^2/2},$$

$$\text{and thus,} \quad \log(1+\lambda y) - \lambda y \overset{(i)}{\geq} \frac{y^2}{m^2}\left(\log(1-m\lambda) + m\lambda\right).$$

Above, (i) can be quickly verified for the case when $\lambda y = 0$, and follows from (37) and (38) otherwise. Rearranging terms, we obtain the first half of the desired result,

$$\log(1+\lambda y) \geq \lambda y + \frac{y^2}{m^2}(\log(1-m\lambda) + m\lambda). \tag{39}$$

Now, for any $y \leq 1 - m$ and $\lambda \in (-1/(1-m), 0]$, we have

$$\lambda y \geq (1-m)\lambda > -1,$$

and proceed similarly to before to obtain

$$\log(1+\lambda y) \geq \lambda y + \frac{y^2}{(1-m)^2}(\log(1+(1-m)\lambda) - (1-m)\lambda),$$

which completes the proof. □

A.7. Proof of Proposition 5

Since sublevel sets of convex functions are convex, it suffices to prove that with probability one, $\mathcal{K}_n^{\mathrm{hgKelly}}(m)$ is a convex function in m on the interval $[0,1]$.

We proceed in three steps. First, we show that if two functions are (a) both nonincreasing (or both nondecreasing), (b) nonnegative, and (c) convex, then their product is convex. Second, we use Step 1 and an induction argument to prove that $\prod_{i=1}^{t}(1+\gamma(X_i/m-1))$ is convex for any fixed $\gamma \in [0,1]$. Third and finally, we show that $\mathcal{K}_n^{\mathrm{hgKelly}}(m)$ is a convex combination of convex functions and is thus itself convex.

Step 1. The claim is that if two functions f and g are (a) both nonincreasing (or both nondecreasing), (b) nonnegative, and (c) convex on a set $\mathcal{S} \subseteq \mathbb{R}$, then their product is also convex on $\mathcal{S}$. Let $x_1, x_2 \in \mathcal{S}$, and let $t \in [0,1]$. Furthermore, abbreviate $f(x_1)$ by f_1, $g(x_1)$ by g_1, and similarly for f_2 and g_2. Writing out the product fg evaluated at $tx_1 + (1-t)x_2$,

$$\begin{aligned}(fg)(tx_1+(1-t)x_2) &= f(tx_1+(1-t)x_2)g(tx_1+(1-t)x_2)\\ &= |f(tx_1+(1-t)x_2)||g(tx_1+(1-t)x_2)|\\ &\leq |tf_2+(1-t)f_2||tg_1+(1-t)g_2|\\ &= t^2 f_1 g_1 + t(1-t)\left(f_1 g_2 + f_2 g_1\right) + (1-t)^2 f_2 g_2,\end{aligned}$$

where the second equality follows from assumption that f and g are nonnegative, and the inequality follows from the assumption that they are both convex. To show convexity of (fg), it then suffices to show that,

$$\Big(tf_1g_1 + (1-t)f_2g_2\Big) - \Big(t^2 f_1 g_1 + t(1-t)\left[f_1 g_2 + f_2 g_1\right] + (1-t)^2 f_2 g_2\Big) \geq 0. \quad (40)$$

To this end, write out the above expression and group terms,

$$\begin{aligned}&\Big(tf_1g_1 + (1-t)f_2g_2\Big) - \Big(t^2 f_1 g_1 + t(1-t)\left[f_1 g_2 + f_2 g_1\right] + (1-t)^2 f_2 g_2\Big)\\ &= (1-t)tf_1g_1 + t(1-t)f_2g_2 - t(1-t)[f_1g_2 + f_2g_1]\\ &= t(1-t)\Big(f_1g_1 + f_2g_2 - f_1g_2 - f_2g_1\Big)\\ &= t(1-t)(f_1-f_2)(g_1-g_2).\end{aligned}$$

Now, notice that $t(1-t) \geq 0$ since $t \in [0,1]$ and that $(f_1-f_2)(g_1-g_2) \geq 0$ by the assumption that f and g are both nonincreasing or nondecreasing. Therefore, we have satisfied the inequality in (40), and thus fg is convex on $\mathcal{S}$.

Step 2. Now, we prove convexity of $\prod_{i=1}^{t}(1+\gamma(X_i/m-1))$ for a fixed $\gamma \in [0,1]$. First note that for any $\gamma \in [0,1]$, $1+\gamma(X_i/m-1)$ is a nonincreasing, nonnegative, and convex function in $m \in [0,1]$. Suppose for the sake of induction that conditions (a), (b), and (c) hold for $\prod_{i=1}^{n-1}(1+\gamma(X_i/m-1))$. By the inductive hypothesis, we

have that

$$\prod_{i=1}^{n}(1+\gamma(X_i/m-1)) = (1+\gamma(X_n/m-1)) \cdot \prod_{i=1}^{n-1}(1+\gamma(X_i/m-1))$$

is a product of functions satisfying (a) through (c). By Step 1, $\prod_{i=1}^{n}(1+\gamma(X_i/m-1))$ is convex in $m \in [0,1]$. A similar argument can be made for $\mathcal{K}_n^-(m)$, but instead of the multiplicands being nonincreasing, they are now nondecreasing.

Step 3. Now, notice that for the evenly-spaced points $(\lambda^{1+}, \ldots, \lambda^{G+})$ on $[0, 1/m]$, we have that $(\gamma^{1+}, \ldots, \gamma^{G+}) = (m\lambda^{1+}, \ldots, m\lambda^{G+})$ are G evenly-spaced points on $[0,1]$. It then follows that for any m and any $g \in \{0, 1, \ldots, G\}$,

$$m \mapsto \prod_{i=1}^{n}(1+\lambda^{g+}(X_i-m))$$

is a nonincreasing, nonnegative, and convex function in $m \in [0,1]$. It follows that

$$\frac{1}{G}\sum_{g=1}^{G}\prod_{i=1}^{n}(1+\lambda^{g+}(X_i-m))$$

is convex in $m \in [0,1]$. A similar argument goes through for $\frac{1}{G}\sum_{g=1}^{G}\prod_{i=1}^{n}(1+\lambda^{g+}(X_i-m)$. Finally, since $\theta \in [0,1]$, we have that

$$\frac{\theta}{G}\sum_{g=1}^{G}\prod_{i=1}^{n}(1+\lambda^{g+}(X_i-m)) + \frac{1-\theta}{G}\sum_{g=1}^{G}\prod_{i=1}^{n}(1+\lambda^{g-}(X_i-m))$$

is a convex combination of convex functions in $m \in [0,1]$. It then follows that

$$\{m \in [0,1] : \mathcal{K}_t^{\mathrm{hgKelly}}(m) < 1/\alpha\}$$

is an interval, which completes the proof. □

A.8. Proof of Proposition 4

Proof of (1) $\Longrightarrow$ (2). By definition of $\mathcal{K}_t^{\mathrm{WoR}}(\mu)$, we have

$$\begin{aligned}
\mathbb{E}\left(\mathcal{K}_t^{\mathrm{WoR}}(\mu) \mid \mathcal{F}_{t-1}\right) &= \prod_{i=1}^{t-1}\left(1+\lambda_i(\mu)\cdot(X_i-\mu_t^{\mathrm{WoR}})\right)\cdot\mathbb{E}\left(1+\lambda_t(\mu)\cdot(X_t-\mu_t^{\mathrm{WoR}}) \mid \mathcal{F}_{t-1}\right) \\
&= \mathcal{K}_{t-1}^{\mathrm{WoR}}(\mu)\cdot\left(1+\lambda_t(\mu)\cdot(\mathbb{E}(X_t \mid \mathcal{F}_{t-1})-\mu_t^{\mathrm{WoR}})\right) \\
&= \mathcal{K}_{t-1}^{\mathrm{WoR}}(\mu).
\end{aligned}$$

Since $\mathcal{K}_0^{\mathrm{WoR}}(\mu) \equiv 1$ by convention, we have that $\mathcal{K}_t^{\mathrm{WoR}}(\mu)$ is a martingale.

Now, note that since $X_t \in [0,1]$ and $\lambda_t^{\mathrm{WoR}}(\mu) \in [-1/(1-\mu_t^{\mathrm{WoR}}), 1/\mu_t^{\mathrm{WoR}}]$ for each t by assumption, we have that $1 + \lambda_t(\mu) \cdot \left(X_t - \mu_t^{\mathrm{WoR}}\right) \geq 0$ and thus $\mathcal{K}_t^{\mathrm{WoR}}(\mu) \geq 0$. Therefore, $\mathcal{K}_t^{\mathrm{WoR}}(\mu)$ is a test martingale.

Proof of (2) $\Longrightarrow$ (1). Suppose that $\mathcal{K}_t^{\mathrm{WoR}}(\mu)$ is a test martingale for any $(\lambda_t(\mu))_{t=1}^N$ with $\lambda_t(\mu) \in [-1/(1-\mu_t^{\mathrm{WoR}}, 1/\mu_t^{\mathrm{WoR}}]$, but suppose for the sake of contradiction that $\mathbb{E}(X_{t^\star} \mid \mathcal{F}_{t^\star-1}) \neq \mu_{t^\star}^{\mathrm{WoR}}$ for some $t^\star \in \{1, 2, \dots\}$. Set $\lambda_1 = \lambda_2 = \cdots = \lambda_{t^\star-1} = 0$ and $\lambda_{t^\star} = 1$. Then,

$$\mathcal{K}_{t^\star}^{\mathrm{WoR}}(\mu) \equiv \mathcal{K}_{t^\star-1}^{\mathrm{WoR}}(\mu) \cdot (1 + \lambda_{t^\star}(X_{t^\star} - \mu_{t^\star}^{\mathrm{WoR}})) = 1 + X_{t^\star} - \mu_{t^\star}^{\mathrm{WoR}}.$$

By assumption of $\mathcal{K}_t^{\mathrm{WoR}}(\mu)$ forming a martingale, we have that $\mathbb{E}\left(\mathcal{K}_{t^\star}^{\mathrm{WoR}}(\mu) \mid \mathcal{F}_{t^\star-1}\right) = \mathcal{K}_{t^\star-1}^{\mathrm{WoR}}(\mu) = 1$. On the other hand, since $\mathbb{E}\left(X_{t^\star} \mid \mathcal{F}_{t^\star-1}\right) \neq \mu_{t^\star}^{\mathrm{WoR}}$, we have

$$\mathbb{E}\left(\mathcal{K}_{t^\star}^{\mathrm{WoR}}(\mu) \mid \mathcal{F}_{t^\star-1}\right) = \mathbb{E}\left(1 + X_{t^\star} - \mu_{t^\star}^{\mathrm{WoR}} \mid \mathcal{F}_{t^\star-1}\right) \neq 1,$$

a contradiction. Therefore, we must have that $\mathbb{E}\left(X_t \mid \mathcal{F}_{t-1}\right) = \mu_t^{\mathrm{WoR}}$ for each t, which completes the proof of (2) $\Longrightarrow$ (1) and Proposition 4. □

A.9. Proof of Theorem 4

The proof that $\mathfrak{B}_t^{\pm,\mathrm{WoR}}$ forms a $(1-\alpha)$-CS for μ proceeds in exactly the same manner as Theorem 3, noting that $\mathbb{E}\left(X_t \mid \mathcal{F}_{t-1}\right) = \mu_t^{\mathrm{WoR}}$ instead of μ.

To show that $\mathfrak{B}_t^{\pm,\mathrm{WoR}}$ is indeed an interval for each $t \geq 1$, we note that the proof of Theorem 3 applies since m_t^{WoR} is increasing or decreasing if and only if m is increasing or decreasing, respectively. □

B. How to bet: deriving adaptive betting strategies

In Section 4.4, we presented CSs and CIs via the hedged capital process. We suggested a specific betting scheme which has strong empirical performance but did not discuss where it came from. In this section, we derive various betting strategies and discuss their statistical and computational properties.

B.1. Predictable plug-ins yield good betting strategies

First and foremost, we will examine why any predictable plug-in for empirical Bernstein-type CSs and CIs (i.e. those recommended in Theorem 2 and Remark 1) yield effective betting strategies. Consider the hedged capital process

$$\begin{aligned}\mathcal{K}_t^{\pm}(m) &:= \max\left\{\theta \prod_{i=1}^{t}(1 + \lambda_i^+(X_i - m)), (1-\theta)\prod_{i=1}^{t}(1 - \lambda_i^-(X_i - m))\right\} \\ &\equiv \max\left\{\theta\mathcal{K}_t^+(m), (1-\theta)\mathcal{K}_t^-(m)\right\}\ ,\end{aligned}$$

where $(\lambda_t^+(m))_{t=1}^\infty$ and $(\lambda_t^-(m))_{t=1}^\infty$ are $[0, 1/m]$-valued and $[0, 1/(1-m)]$-valued predictable sequences as in Theorem 3. First, consider the "positive" capital process, $\mathcal{K}_t^+(\mu)$ evaluated at $m = \mu$. An inequality that has been repeatedly used to derive

empirical Bernstein inequalities (Howard et al., 2020, 2021; Waudby-Smith and Ramdas, 2020), including the current paper is the following due to Fan et al. (2015, equation 4.12): for any $y \geq -1$ and $\lambda \in [0,1)$, we have

$$\log(1+\lambda y) \geq \lambda y - 4\psi_E(\lambda) y^2. \tag{41}$$

where $\psi_E(\lambda)$ is as defined in (14). If the predictable sequence $(\lambda_t^+(m))_{t=1}^\infty$ is further restricted to $[0,1)$, then by (41) we have

$$\begin{aligned}
\mathcal{K}_t^+(\mu) := \prod_{i=1}^t (1+\lambda_i^+(X_i-\mu)) \ &\geq \exp\left(\sum_{i=1}^t \lambda_i^+(X_i-\mu) - \sum_{i=1}^t 4(X_i-\mu)^2\psi_E(\lambda_i^+)\right) \\
&\overset{(i)}{\approx} \exp\left(\sum_{i=1}^t \lambda_i^+(X_i-\mu) - \sum_{i=1}^t 4(X_i-\widehat{\mu}_{i-1})^2\psi_E(\lambda_i^+)\right) \\
&= M_t^{\text{PrPl-EB}}(\mu),
\end{aligned}$$

where (i) follows from the approximations $\widehat{\mu}_{t-1} \approx \mu$ for large t. Not only does the approximate inequality $\mathcal{K}_t^+(\mu) \gtrsim M_t^{\text{PrPl-EB}}(\mu)$ shed light on why a sensible empirical Bernstein predictable plug-in translates to a sensible betting strategy, but also why we might expect $\mathcal{K}_t^+(m)$ to be more powerful than $M_t^{\text{PrPl-EB}}(m)$ for the same $[0,1)$-valued predictable sequence $(\lambda_t^+(m))_{t=1}^\infty$. Moreover, $\mathcal{K}_t^+(m)$ has the added flexibility of allowing $(\lambda_t(m))_{t=1}^\infty$ to take values in $[0,1/m] \supset [0,1)$ which we find — through simulations — tends to improves empirical performance (see Figure 19 in Section E.2.2). Finally, a similar story holds for $\mathcal{K}_t^-(\mu)$ with the added caveat that $(\lambda_t^-)_{t=1}^\infty$ can instead take values in $[0,1/(1-m)] \supset [0,1)$ which as before, seems to improve empirical performance.

Despite the success of predictable plug-ins as betting strategies, it is natural to wonder whether it is preferable to focus on directly maximizing capital over time. As will be seen in the following section, these capital-maximizing approaches tend to have improved empirical performance, but are not always guaranteed to produce convex confidence sets (i.e. intervals). Nevertheless, it is worth examining some of these strategies both for their intuitive appeal and excellent empirical performance.

B.2. Growth rate adaptive to the particular alternative (GRAPA)

As alluded to in Section 6, Kelly Jr (1956) dealt with capital processes, betting strategies, etc. in the fields of information and communication theory in the pursuit of maximizing the information rate over a channel. Kelly suggested that an effective betting strategy is one that maximizes a gambler's expected *log-capital* — i.e. the growth rate of the gambler's capital — under a particular alternative.§ However, Kelly's setup was a simplified special case of ours: Kelly's observations were binary, and the exact alternative was assumed known, while ours are merely bounded in

§This objective has also been arrived at indirectly as the dual in optimization programs for deriving regret bounds for Kullback-Leibler-based UCB algorithms in multi-armed bandit problems (Honda and Takemura, 2010; Cappé et al., 2013).

$[0,1]$ with an unknown alternative. Nevertheless, the principle of maximizing the log-capital can be adapted to our setting under bounded observations and an unknown alternative. We summarize this adaptation here and refer to it as maximizing the "growth rate adaptive to the particular alternative" or "GRAPA" for short.

Write the log-capital process at time t as

$$\ell_t(\lambda_1^t, m) := \log(\mathcal{K}_t(m)) = \sum_{i=1}^{t} \log(1 + \lambda_i(m)(X_i - m)), \tag{42}$$

for a general $[-1/(1-m), 1/m]$-valued sequence $(\lambda_t(m))_{t=1}^{\infty}$. If we were to choose a single value of $\lambda^{\mathrm{HS}} := \lambda_1 = \cdots = \lambda_t$ which maximizes the log-capital ℓ_t "in hindsight" (i.e. based on *all* of the previous data), then this value is given by

$$\frac{\partial \ell_t(\lambda^{\mathrm{HS}}, m)}{\partial \lambda^{\mathrm{HS}}} = \sum_{i=1}^{t} \frac{X_i - m}{1 + \lambda^{\mathrm{HS}}(X_i - m)} \overset{\text{set}}{=} 0.$$

However, λ^{HS} is clearly not predictable. Following Kumon et al. (2011) (who referred to this as the "sequential optimization strategy"), we set $(\lambda_t^{\mathrm{GRAPA}}(m))_{t=1}^{\infty}$ such that

$$\frac{1}{t-1} \sum_{i=1}^{t-1} \frac{X_i - m}{1 + \lambda_t^{\mathrm{GRAPA}}(m)(X_i - m)} \overset{\text{set}}{=} 0, \tag{43}$$

truncated to lie between $(-c/(1-m), c/m)$ using some $c \leq 1$. Importantly, $\lambda_t^{\mathrm{GRAPA}}(m)$ only depends on $X_1, \ldots, X_{t-1}$, and is thus predictable.

This rule is a sequentially adaptive version of the worst-case "GROW" criterion of Grünwald et al. (2019). To see the connection, one can derive (43) from a slightly different motivation. At the t-th step, we want to choose $\lambda_t(m)$ so that the wealth multiplier $(1 + \lambda_t(m)(X_t - m))$ is as large as possible. The ideal choice would be

$$\lambda_t^*(m) := \operatorname*{argmax}_{\lambda \in [-1/(1-m), 1/m]} \mathbb{E}_{P^\mu}[\log(1 + \lambda(X_t - m)) \mid \mathcal{F}_{t-1}], \tag{44}$$

where P^μ is the unknown true distribution. Writing down the stationary condition for this optimization problem by differentiating through the expectation, we get

$$\mathbb{E}_{P^\mu}\left[\frac{X_t - m}{1 + \lambda_t^*(m)(X_t - m)} \mid \mathcal{F}_{t-1}\right] = 0. \tag{45}$$

Since P^μ is unknown, using a simple empirical plug-in estimator yields (43).

CSs constructed from $(\lambda_t^{\mathrm{GRAPA}}(m))_{t=1}^{\infty}$ tend to have excellent empirical performance, but can be prohibitively slow due to the required root-finding in (43) for each time t and $m \in [0,1]$ (or a sufficiently fine grid of $[0,1]$). A similar but computationally inexpensive alternative to GRAPA is "approximate GRAPA" (aGRAPA), which we derive now.

B.3. Approximate GRAPA (aGRAPA)

Rather than solve (43), we take the Taylor approximation of $(1+y)^{-1}$ by $(1-y)$ for $y \approx 0$ to obtain

$$\begin{aligned}\frac{1}{t-1}\sum_{i=1}^{t-1}\frac{X_i-m}{1+\lambda_t^{\mathrm{aGRAPA}}(m)(X_i-m)} &\approx \frac{1}{t-1}\sum_{i=1}^{t-1}\left(1-\lambda_t^{\mathrm{aGRAPA}}(m)(X_i-m)\right)(X_i-m)\\ &= \frac{1}{t-1}\sum_{i=1}^{t-1}(X_i-m) - \frac{\lambda_t^{\mathrm{aGRAPA}}(m)}{t-1}\sum_{i=1}^{t-1}(X_i-m)^2\\ &\overset{\text{set}}{=} 0,\end{aligned}$$

which, after appropriate truncation leads what we call the "approximate GRAPA" (aGRAPA) betting strategy,

$$\lambda_t^{\mathrm{aGRAPA}}(m) := -\frac{c}{1-m} \vee \frac{\widehat{\mu}_{t-1}-m}{\widehat{\sigma}_{t-1}^2+(\widehat{\mu}_{t-1}-m)^2} \wedge \frac{c}{m},$$

for some truncation level $c \leq 1$. This expression is quite natural: we bet more aggressively if our empirical mean is far away from m, and are further emboldened if the empirical variance is small.

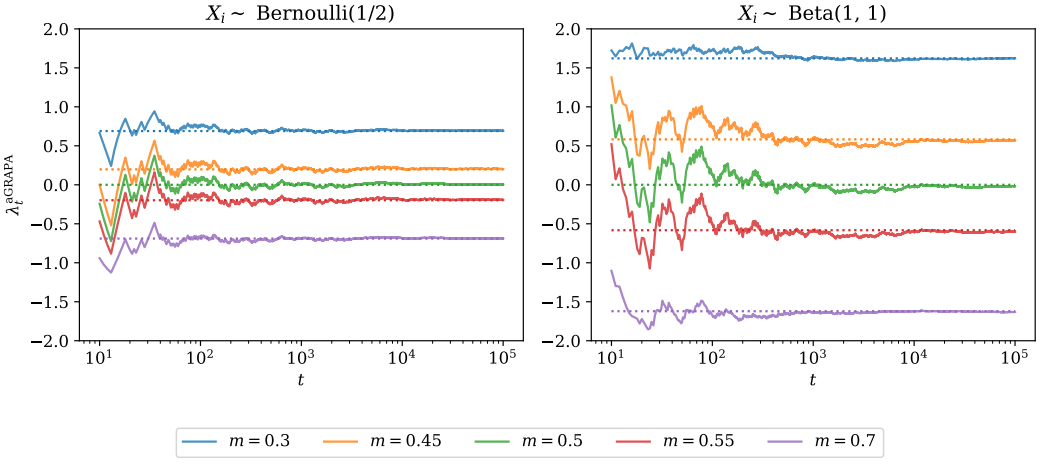


Figure 10. $\lambda_t^{\mathrm{aGRAPA}}$ for various values of m under two distributions: Bernoulli(1/2) and Beta(1, 1). The dotted lines show the "oracle" bets, meaning $\lambda_t^{\mathrm{aGRAPA}}$ with estimates of the mean and variance replaced by their true values. As time passes, bets stabilize and approach their oracle quantities.

As alluded to at the end of Section B.1, CSs derived using the capital process $\mathcal{K}_t(m)$ with arbitrary betting schemes are not always guaranteed to produce a convex set (interval). In fact, it is possible to construct scenarios where the sublevel sets of $\mathcal{K}_t^{\mathrm{aGRAPA}}(m)$ are nonconvex in m (see Section E.4 for an example). In our experience, this type of situation is not common, and one must actively search for such pathological examples.

B.4. Lower-bound on the wealth (LBOW)

Instead of maximizing $\log(\mathcal{K}_t(m))$, we may aim to do so for a tight lower-bound on the wealth (LBOW). This technique has proven useful in the game-theoretic probability literature (Shafer and Vovk, 2001, Proof of Lemma 3.3) and (Cutkosky and Orabona, 2018, Proof of Theorem 1). Our lower bound will rely on an extension of Fan's inequality (41) to $\lambda \in (-1/(1-m), 1/m)$, summarized in the following lemma.

LEMMA 3. *If $y \geq -m$, then for any $\lambda \in [0, 1/m)$, we have*

$$\log(1+\lambda y) \geq \lambda y + \frac{y^2}{m^2}(\log(1-m\lambda) + m\lambda).$$

On the other hand, if $y \leq 1-m$, then for any $\lambda \in (-1/(1-m), 0]$, we have

$$\log(1+\lambda y) \geq \lambda y + \frac{y^2}{(1-m)^2}(\log(1+(1-m)\lambda) - (1-m)\lambda).$$

Thus, for $y \in [-m, 1-m]$, both of the above inequalities hold.

The proof is an easy generalization of inequality (41) by Fan et al. (2015), and also follows from similar observations about the subexponential function ψ_E in Howard et al. (2020, 2021), but we prove it from first principles in Section A.6 for completeness. Using Lemma 3, we have for $\lambda^{\mathrm{L}+} \in [0, 1/m)$, the following lower-bound on $\ell(\lambda^{\mathrm{L}+}, m)$,

$$\begin{aligned}\ell(\lambda^{\mathrm{L}+}, m) &:= \log\left(\prod_{i=1}^{t}(1+\lambda^{\mathrm{L}+}(X_i - m))\right) \\ &\geq \lambda^{\mathrm{L}+}\sum_{i=1}^{t}(X_i - m) + \frac{\log(1-m\lambda^{\mathrm{L}+}) + m\lambda^{\mathrm{L}+}}{m^2}\sum_{i=1}^{t}(X_i-m)^2, \qquad (46)\end{aligned}$$

and for $\lambda^{\mathrm{L}-} \in (-1/(1-m), 0]$, we have

$$\begin{aligned}\ell(\lambda^{\mathrm{L}-}, m) &:= \log\left(\prod_{i=1}^{t}(1+\lambda^{\mathrm{L}-}(X_i - m))\right) \\ &\geq \lambda^{\mathrm{L}-}\sum_{i=1}^{t}(X_i - m) + \frac{\log(1+(1-m)\lambda^{\mathrm{L}-}) - (1-m)\lambda^{\mathrm{L}-}}{(1-m)^2}\sum_{i=1}^{t}(X_i-m)^2. \qquad (47)\end{aligned}$$

Importantly, if $\sum_{i=1}^{t}(X_i - m)$ is positive, then (46) is concave, while if negative, (47) is concave. Maximizing (46) or (47) depending on the sign of $\sum_{i=1}^{t}(X_i - m)$ we obtain the following "hindsight" choice for λ^{L},

$$\lambda^{\mathrm{L}} = \begin{cases} \dfrac{\sum_{i=1}^{t}(X_i-m)}{m\sum_{i=1}^{t}(X_i-m) + \sum_{i=1}^{t}(X_i-m)^2} & \text{if } \displaystyle\sum_{i=1}^{t}(X_i-m) \geq 0, \\ \dfrac{\sum_{i=1}^{t}(X_i-m)}{-(1-m)\sum_{i=1}^{t}(X_i-m) + \sum_{i=1}^{t}(X_i-m)^2} & \text{if } \displaystyle\sum_{i=1}^{t}(X_i-m) \leq 0. \end{cases}$$

Of course, this choice of λ^{L} is not predictable and thus is not a valid betting strategy in the framework of the current paper. This motivates the following strategy, $(\lambda_t^{\mathrm{L}}(m))_{t=1}^{\infty}$ given by

$$\lambda_t^{\mathrm{L}}(m) := \frac{-c}{1-m} \vee \frac{\widehat{\mu}_{t-1} - m}{\omega_{t-1}|\widehat{\mu}_{t-1} - m| + \widehat{\sigma}_{t-1}^2 + (\widehat{\mu}_{t-1} - m)^2} \wedge \frac{c}{m}, \tag{48}$$

$$\text{where} \quad \omega_t := \begin{cases} m & \text{if } \widehat{\mu}_t - m \geq 0 \ , \\ 1-m & \text{if } \widehat{\mu}_t - m < 0 \ . \end{cases}$$

Similarly to the aGRAPA betting procedure, LBOW is computationally-inexpensive but is not guaranteed to produce an interval. The expression also carries similar intuition to the GRAPA case.

B.5. Online Newton Step (ONS-m)

Betting algorithms play an essential role in online learning as several optimization problems can be framed in terms of coin-betting games (Cutkosky and Orabona, 2018; Orabona and Tommasi, 2017; Jun et al., 2017; Jun and Orabona, 2019). While the downstream application is different, the game-theoretic techniques of maximizing wealth are almost immediately applicable to the problem at hand. Here, we consider a slight modification to the Online Newton Step (ONS) algorithm due to Cutkosky and Orabona (2018).

Algorithm 1: ONS-m.

Result: $(\lambda_t^{\mathrm{O}}(m))_{t=1}^{T}$
$\lambda_1^{\mathrm{O}}(m) \leftarrow 1$;
for $t \in \{1, \ldots, T-1\}$ **do**
  $y_t \leftarrow X_t - m$;
  Set $z_t \leftarrow y_t/(1 - y_t\lambda_t^{\mathrm{O}}(m))$;
  $A_t \leftarrow 1 + \sum_{i=1}^{t} z_i^2$;
  $\lambda_{t+1}^{\mathrm{O}}(m) \leftarrow \frac{-c}{1-m} \vee \left(\lambda_t^{\mathrm{O}}(m) - \frac{2}{2-\log(3)}\frac{z_t}{A_t}\right) \wedge \frac{c}{m}$;
end

Through simulations, we find that ONS-m performs competitively. However, its lack of closed-form expression makes it a slightly more computationally-expensive alternative to aGRAPA and LBOW, but not nearly as expensive as GRAPA (see Table 2).

B.6. Diversified Kelly betting (dKelly)

Instead of committing to one betting strategy such as aGRAPA or LBOW, we can simply take the average capital among D separate strategies. This follows from the fact that an average of test martingales is itself a test martingale. That is, if

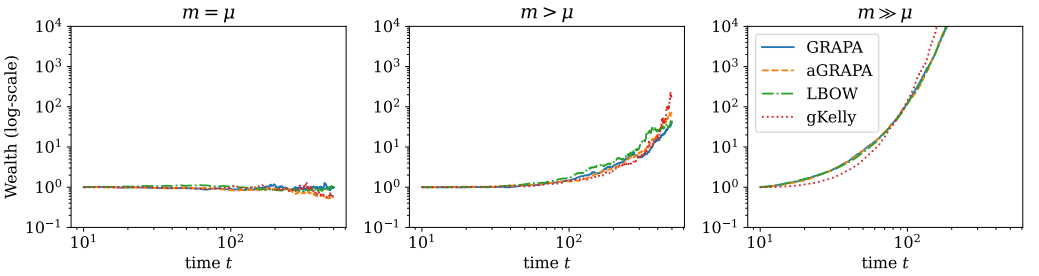


Figure 11. Comparison of the wealth process under various game-theoretic betting strategies with 100 repeats. In this example, the 1000 observations are drawn from a Beta(10, 10) distribution, and the candidate means m being tested are 0.5, 0.51, and 0.55 (from left to right). Notice that these strategies perform similarly, but have varying computational costs (see Table 2).

$(\lambda_t^1)_{t=1}^\infty, (\lambda_t^2)_{t=1}^\infty, \ldots, (\lambda_t^D)_{t=1}^\infty$ are D separate betting strategies, then

$$\mathcal{K}_t^{\mathrm{dKelly}}(\mu) := \frac{1}{D}\sum_{d=1}^{D}\prod_{i=1}^{t}\left(1+\lambda_i^d(\mu)(X_i-\mu)\right)$$

forms a test martingale. Following Kelly's original motivation to maximize (expected) log-capital, notice that by Jensen's inequality,

$$\log\left(\mathcal{K}_t^{\mathrm{dKelly}}\right) > \frac{1}{D}\sum_{d=1}^{D}\log\left(\prod_{i=1}^{t}\left(1+\lambda_i^d(\mu)(X_i-\mu)\right)\right).$$

In other words, the log-capital of the diversified bets is strictly larger than the average log-capital among the diverse candidate bets.

Grid Kelly betting (gKelly). While it is possible to use any finite collection of strategies, we focus our attention on a particularly simple (and useful) example where the bets are constant values on a grid. Specifically, divide the interval $[-1/(1-m), 1/m]$ up into G evenly-spaced points $\lambda^1, \ldots, \lambda^G$. Then define the gKelly capital process $\mathcal{K}_t^{\mathrm{gKelly}}$ by

$$\mathcal{K}_t^{\mathrm{gKelly}}(m) := \frac{1}{G}\sum_{g=1}^{G}\prod_{i=1}^{t}(1+\lambda^g(X_i-m)).$$

When used to construct confidence sequences for μ, $\mathcal{K}_t^{\mathrm{gKelly}}$ demonstrates excellent empirical performance. Moreover, this procedure can be slightly modified into "Hedged gKelly" (hgKelly) so that confidence sequences constructed using gKelly are intervals almost surely.

In order to mimic the unknown optimal λ^*, D or G should not be kept constant, but itself grow slowly (say logarithmically) with t. In game-theoretic terms, one

should slowly add more strategies to the portfolio, in order to asymptotically match the performance of the optimal one over time. (When adding a new λ^g to an existing mixture, it obviously only begins to contribute to the wealth from the following step onwards; formally G would be replaced by G_t, and $\prod_{i=1}^t(1+\lambda^g(X_i-m)$ would be replaced by $\prod_{i=t_g}^t(1+\lambda^g(X_i-m)$ if λ^g was first introduced after t_g-1 steps.)

Hedged gKelly. First, divide the interval $[-1/(1-m),0]$ and $[0,1/m]$ into G evenly-spaced points: $(\lambda^{1-},\ldots,\lambda^{G-})$ and $(\lambda^{1+},\ldots,\lambda^{G+})$, respectively. Then define the "Hedged grid Kelly capital process" $\mathcal{K}_t^{\mathrm{hgKelly}}$ given by

$$\mathcal{K}_t^{\mathrm{hgKelly}}(m) := \frac{\theta}{G}\sum_{g=1}^{G}\prod_{i=1}^{t}\left(1+\lambda^{g+}(X_i-m)\right) + \frac{1-\theta}{G}\sum_{g=1}^{G}\prod_{i=1}^{t}\left(1+\lambda^{g-}(X_i-m)\right),$$

where $\theta\in[0,1]$ (a reasonable default being $\theta=1/2$).

PROPOSITION 5. *If $(X_t)_{t=1}^\infty \sim P$ for some $P\in\mathcal{P}^\mu$, then $\mathcal{K}_t^{\mathrm{hgKelly}}(\mu)$ forms a test martingale and $\mathfrak{B}_t^{\mathrm{hgKelly}} := \left\{m\in[0,1]: \mathcal{K}_t^{\mathrm{hgKelly}}(m) < 1/\alpha\right\}$ is a CS for μ that forms an interval for each $t\geq 1$.*

The proof in Section A.7 proceeds by showing that $\mathcal{K}_t^{\mathrm{hgKelly}}$ is a convex function of m and hence its sublevel sets are intervals.

B.7. Confidence Boundary (ConBo)

The aforementioned strategies benefit from targeting bets against a particular null hypothesis, H_0^m for each $m\in[0,1]$, but this has the drawback of $\mathcal{K}_t(m)$ potentially not being quasiconvex in m. One of the advantages of the hedged capital process as described in Theorem 3 is that $\mathcal{K}_t^{\pm}(m)$ is always quasiconvex, and thus its sublevel sets (and hence the confidence sets $\mathfrak{B}_t^{\pm}$) are intervals.

In an effort to develop game-theoretic betting strategies which generate confidence sets which are intervals, we present the Confidence Boundary (ConBo) bets. Rather than bet against the null hypotheses H_0^m for each $m\in[0,1]$, consider two sequences of nulls, $(H_0^{u_t})_{t=1}^\infty$ and $(H_0^{l_t})_{t=1}^\infty$ corresponding to upper and lower confidence boundaries, respectively. The ConBo bet λ_t^{CB} is then targeted against u_{t-1} and l_{t-1} using *any* game-theoretic betting strategy (e.g. $*$GRAPA, $*$Kelly, LBOW, or ONS-m). Letting $\lambda_t^{\mathrm{G}}(m)$ be any such strategy, we summarize the ConBo betting scheme in Algorithm 2.

COROLLARY 1 (CONFIDENCE BOUNDARY CS **[CONBO]**). *In Algorithm 2,*

$$\mathfrak{B}_t^{\mathrm{CB}} \quad \text{forms a } (1-\alpha)\text{-CS for } \mu,$$

as does $\bigcap_{i\leq t}\mathfrak{B}_i^{\mathrm{CB}}$. Further, $\mathfrak{B}_t^{\mathrm{CB}}$ is an interval for any $t\geq 1$.

We can also adapt the ConBo betting scheme outlined in Algorithm 2 to the without-replacement setting by replacing m by m_t^{WoR} for each time t.

Algorithm 2: ConBo

Result: $(\mathcal{K}_t^{\mathrm{CB}}(m))_{t=1}^T$
$l_0 \leftarrow 0$; $u_0 \leftarrow 1$;
$\mathcal{K}_0^{\mathrm{CB}+}(m) \leftarrow \mathcal{K}_0^{\mathrm{CB}-}(m) \leftarrow 1$;
for $t \in \{1, \ldots, T\}$ **do**
  $\lambda_t^{\mathrm{CB}+} \leftarrow \max\left\{\lambda_t^{\mathrm{G}}(l_{t-1}), 0\right\} \wedge c/m$; `// Compute ConBo bets`
  $\lambda_t^{\mathrm{CB}-} \leftarrow \left|\min\left\{\lambda_t^{\mathrm{G}}(u_{t-1}), 0\right\}\right| \wedge c/(1-m)$;
  $\mathcal{K}_t^{\mathrm{CB}+}(m) \leftarrow \left[1 + \lambda_t^{\mathrm{CB}+}(X_t - m)\right] \cdot \mathcal{K}_{t-1}^{\mathrm{CB}+}(m)$; `// Update capital`
  $\mathcal{K}_t^{\mathrm{CB}-}(m) \leftarrow \left[1 - \lambda_t^{\mathrm{CB}-}(X_t - m)\right] \cdot \mathcal{K}_{t-1}^{\mathrm{CB}-}(m)$;
  $\mathcal{K}_t^{\mathrm{CB}}(m) \leftarrow \max\left\{\theta\mathcal{K}_t^{\mathrm{CB}+}(m), (1-\theta)\mathcal{K}_t^{\mathrm{CB}-}(m)\right\}$; `// Hedging`
  $\mathfrak{B}_t^{\mathrm{CB}} \leftarrow \{m \in [0,1] : \mathcal{K}_t(m) < 1/\alpha\}$;
  $l_t \leftarrow \inf \mathfrak{B}_t^{\mathrm{CB}}$; `// Update confidence boundaries to bet against`
  $u_t \leftarrow \sup \mathfrak{B}_t^{\mathrm{CB}}$;
end

Corollary 2 (WoR confidence boundary CS **[ConBo-WoR]**). *Under the same conditions as Theorem 4, define* $\lambda_t^{\text{CB-WoR}+}$ *and* $\lambda_t^{\text{CB-WoR}-}$ *as in Algorithm 2 but with* m *replaced by* m_t^{WoR}*. Then,*

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

as does $\bigcap_{i \le t} \mathfrak{B}_i^{\text{CB-WoR}}$*. Further,* $\mathfrak{B}_t^{\text{CB-WoR}}$ *is an interval for each* $t \ge 1$*.*

B.8. Sequentially Rebalanced Portfolio (SRP)

Implicitly, none of the aforementioned strategies take advantage of "rebalancing", meaning the ability to take ones capital $\mathcal{K}_t$ at time t, diversify it in any manner at time $t+1$, and repeat. This has had the mathematical advantage of being able to write the resulting capital process $(\mathcal{K}_t(m))_{t=1}^\infty$ in the following general, but closed-form expression:

$$\mathcal{K}_t(m) := \sum_{d=1}^{D} \theta_d \prod_{i=1}^{t} (1 + \lambda_i^d(m) \cdot (X_i - m)),$$

where $D \ge 1$ is as in Section B.6, $(\lambda_t^1(m))_{t=1}^\infty, \ldots, (\lambda_t^D(m))_{t=1}^\infty$ are $[-1/(1-m), 1/m]$-valued predictable sequences as usual, and $(\theta_d)_{d=1}^D$ are convex weights such that $\sum_{d=1}^D \theta_d = 1$. However, a more general capital process martingale can be written but instead of having a closed-form product expression, it can be written recursively as

$$\mathcal{K}_t^{\mathrm{SRP}}(m) := \sum_{d=1}^{D_t} (1 + \lambda_t^d(m) \cdot (X_t - m)) \cdot \theta_t^d \cdot \mathcal{K}_{t-1}^{\mathrm{SRP}}(m), \tag{49}$$

where $(\lambda_t^d)_{d=1}^{D_t}$ are $[1/(1-m), 1/m]$-valued predictable bets, $(\theta_t^d)_{d=1}^{D_t}$ are predictable convex weights that sum to 1 (conditional on X_1^{t-1}), and we have set the initial capital $\mathcal{K}_0^{\mathrm{SRP}}(m)$ to 1 as usual.

Adopting the betting interpretation, (49) is a rather intuitive procedure. At each time step t, the gambler divides their previous capital $\mathcal{K}_{t-1}^{\mathrm{SRP}}(m)$ up into $D_t \geq 1$ portions given by $\theta_t^1 \cdot K_{t-1}^{\mathrm{SRP}}(m), \ldots, \theta_t^{D_t} \cdot \mathcal{K}_{t-1}^{\mathrm{SRP}}(m)$, then invests these wealths with bets $\lambda_t^1(m), \ldots, \lambda_t^{D_t}(m)$, respectively. The gambler's wealths are then updated to

$$(1 + \lambda_t^1(m) \cdot (X_t - m)) \cdot \theta_t^1 \cdot \mathcal{K}_{t-1}^{\mathrm{SRP}}(m), \ldots, (1 + \lambda_t^{D_t}(m) \cdot (X_t - m)) \cdot \theta_t^{D_t} \cdot \mathcal{K}_{t-1}^{\mathrm{SRP}}(m),$$

which are then combined via summation to yield a final capital of (49).

It is now routine to check that the process given by (49) is a nonnegative martingale when evaluated at μ since

$$\begin{aligned}
\mathbb{E}\left(\mathcal{K}_t^{\mathrm{SRP}}(\mu) \mid X_1^{t-1}\right) &= \sum_{d=1}^{D_t} \mathcal{K}_{t-1}^{\mathrm{SRP}}(\mu) \cdot \theta_t^{D_t} \cdot \left(1 + \lambda_t(\mu) \left(\underbrace{\mathbb{E}(X_t \mid X_1^{t-1}) - \mu}_{=0}\right)\right) \\
&= \mathcal{K}_{t-1}^{\mathrm{SRP}}(\mu) \underbrace{\sum_{d=1}^{D_t} \theta_t^{D_t}}_{=1} = \mathcal{K}_{t-1}^{\mathrm{SRP}}(\mu).
\end{aligned}$$

Note that SRP is the most general and customizable betting strategy presented in this paper, since it can be composed of any of the previously discussed strategies, and includes each of them as a special case.

C. Simulations

This section contains a comprehensive set of simulations comparing our new confidence sets presented against previous works. We present simulations for building both time-uniform CSs and fixed-time CIs with or without replacement. Each of these are presented under four distributional "themes": (1) discrete, high-variance; (2) discrete, low-variance; (3) real-valued, evenly spread; and (4) real-valued, concentrated.

C.1. Time-uniform confidence sequences (with replacement)

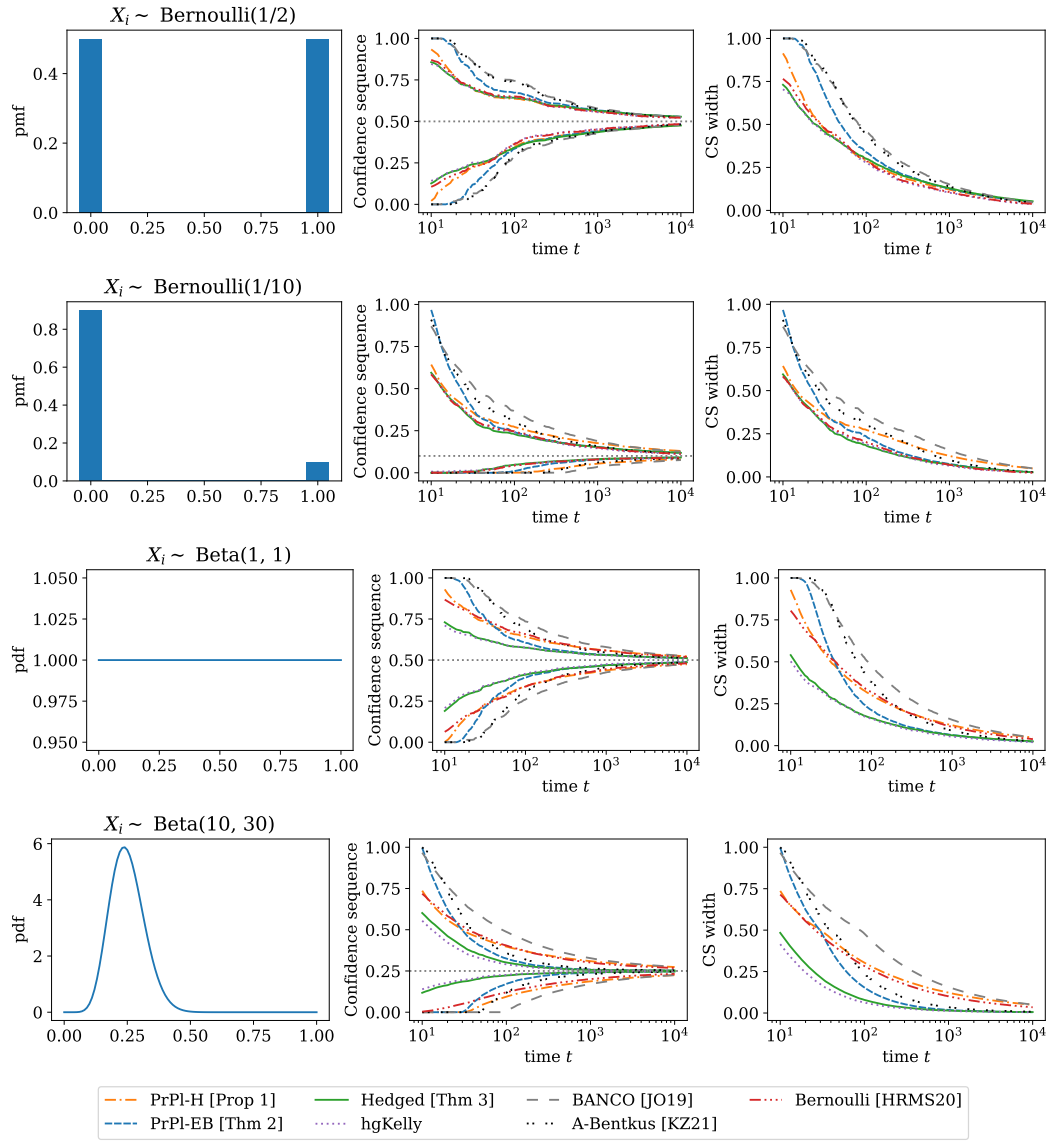


Figure 12. Comparing Hedged, hgKelly, PrPl-EB, and PrPl-H CSs alongside other time-uniform confidence sequences in the literature; further details in Section D.1. Clearly, the betting approach is dominant in all settings.

C.2. Fixed-time confidence intervals (with replacement)

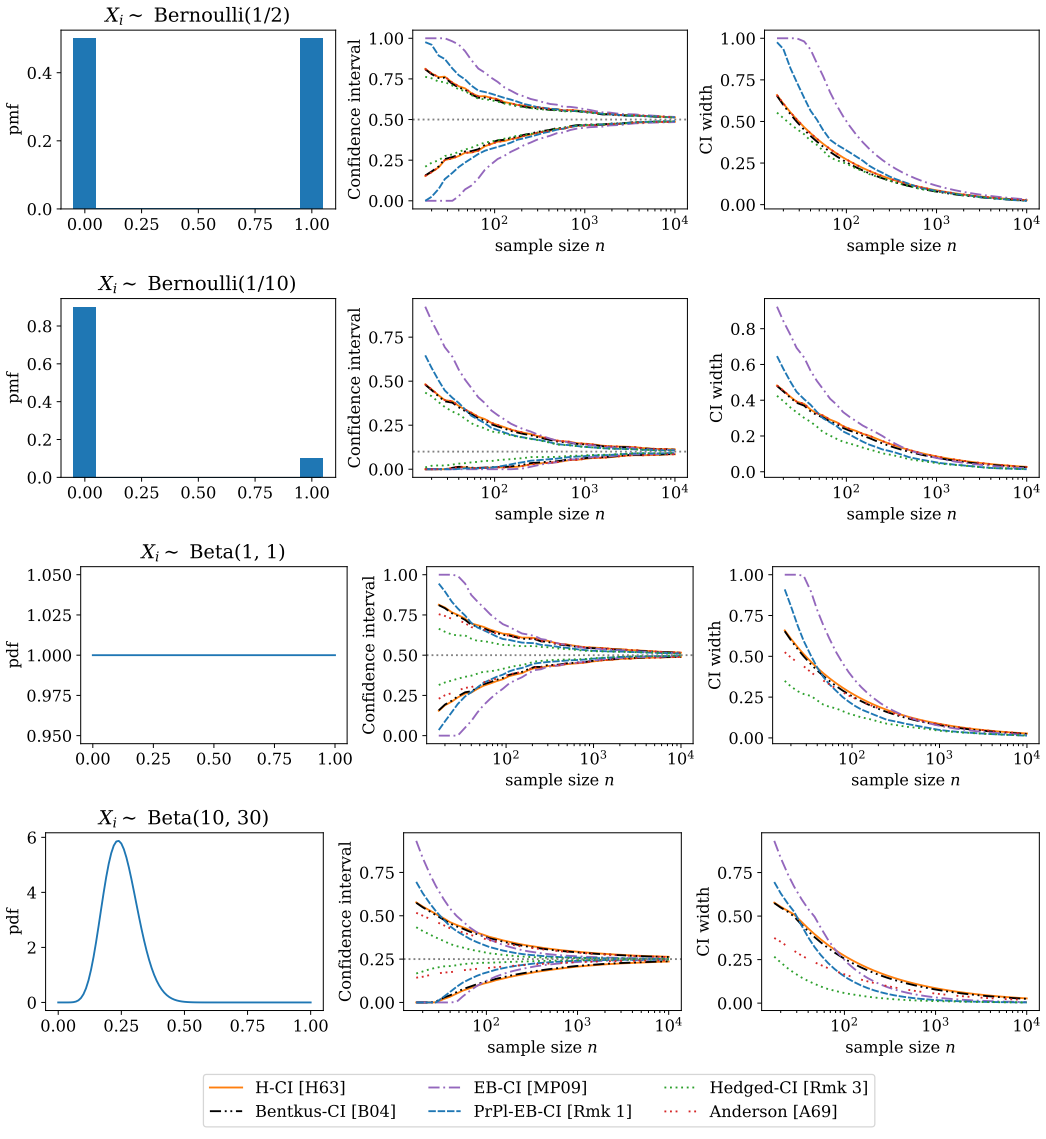


Figure 13. Hedged capital, Anderson, Bentkus, Maurer-Pontil empirical Bernstein, and predictable plug-in empirical Bernstein CIs under four distributional scenarios. Further details can be found in Section D.2. Clearly, the betting approach is dominant in all settings.

C.3. Time-uniform confidence sequences (without replacement)

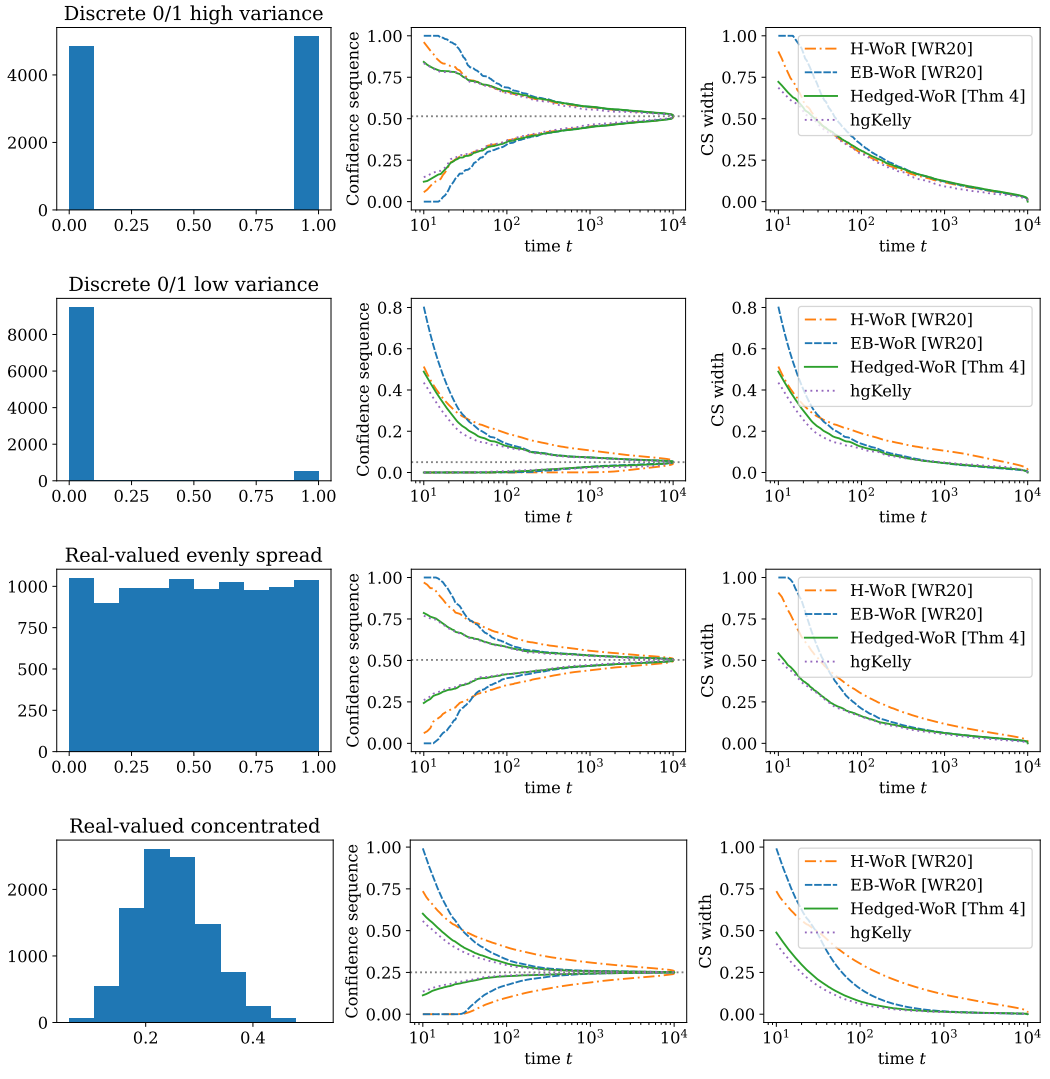


Figure 14. Hedged capital, Hoeffding, and empirical Bernstein CSs for the mean of a finite set of bounded numbers when sampling WoR. Further details can be found in Section D.3. Clearly, the betting approach is dominant in all settings.

C.4. Fixed-time confidence intervals (without replacement)

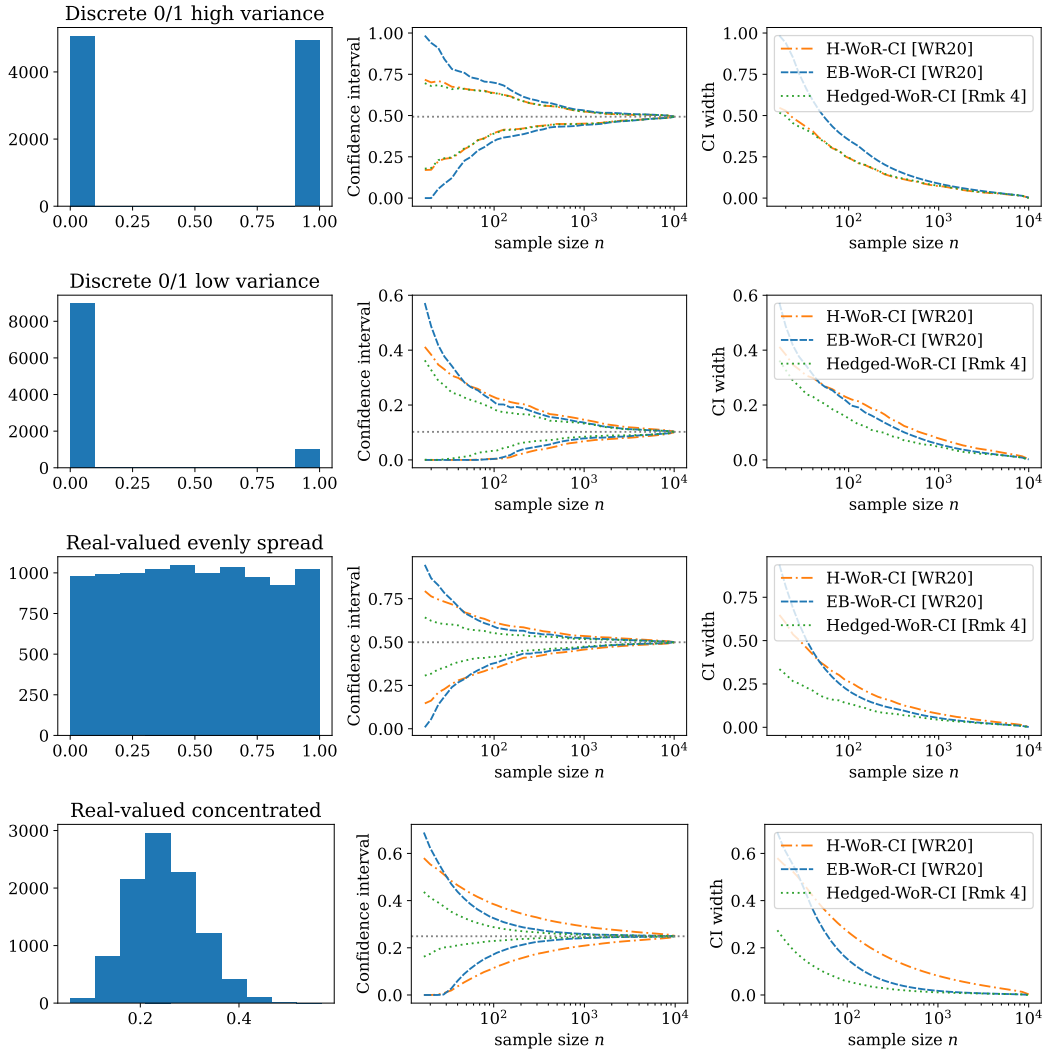


Figure 15. Fixed-time hedged capital, Hoeffding-type, and empirical Bernstein-type CIs for the mean of a finite set of bounded numbers when sampling WoR. Further details can be found in Section D.4. Clearly, the two betting approaches (Hedged and ConBo) are dominant in all settings.

Table 2. Typical computation time for constructing a CS from time 1 to 10^3 for the mean of Bernoulli$(1/2)$-distributed random variables. The three betting CSs were computed for 1000 evenly-spaced values of m in $[0, 1]$, while a coarser grid would have sped up computation. All CSs were calculated on a laptop powered by a quad-core 2GHz 10th generation Intel Core i5. Parallelization was carried out using the Python library, `multiprocess` (McKerns et al., 2011).

Betting scheme	Interval (a.s.)	Computation time (seconds)
ConBo+LBOW	✓	0.08
Hedged+$(\lambda_t^{\mathrm{PrPl}\pm})_{t=1}^{\infty}$	✓	0.25
hgKelly ($G = 20$)	✓	1.38
aGRAPA		0.35
LBOW		0.25
ONS-m		12.45
Kelly		197.38

D. Simulation details

In each simulation containing confidence sequences or intervals and their widths, we took an average over 5 random draws from the relevant distribution. For example, in the "Time-uniform confidence sequences" plot of Figure 1, the CSs (PrPl-H, PrPl-EB, and Hedged) were averaged over 5 random draws from a Beta$(10, 30)$ distribution. Computation times for various strategies are given in Table 2.

D.1. Time-uniform confidence sequences (with replacement)

Each of the CSs considered in the time-uniform (with replacement) case are presented as explicit theorems and propositions throughout the paper. Specifically,

- **PrPl-H**: Predictable plug-in Hoeffding (Proposition 1);
- **PrPl-EB**: Predictable plug-in empirical Bernstein (Theorem 2);
- **Hedged**: Hedged capital process (Theorem 3); and
- **hgKelly**: Hedged grid-Kelly (Proposition 5).

Bernoulli [HRMS20] Section C compared these against the conjugate mixture sub-Bernoulli confidence sequence by Howard et al. (2021), recalled below.

Hoeffding (1963, Equation (3.4)), presented the sub-Bernoulli upper-bound on the moment generating function of bounded random variables for any $\lambda > 0$:

$$\mathbb{E}_P\left(\exp\left\{\lambda(X_i - \mu)\right\}\right) \leq 1 - \mu + \mu \exp\{\lambda\},$$

which can be used to construct an e-value by noting that

$$\mathbb{E}_P\left(\exp\left\{\lambda(X_i - \mu) - \log(1 - \mu + \mu e^{\lambda})\right\} \mid \mathcal{F}_{i-1}\right) \leq 1.$$

Then, Howard et al. (2021) showed that the cumulative product process

$$\prod_{i=1}^{t}\left(\exp\left\{\lambda(X_i - \mu) - \log(1 - \mu + \mu e^{\lambda})\right\}\right) \tag{50}$$

forms a test supermartingale, as does a mixture of (50) for any probability distribution $F(\lambda)$ on $\mathbb{R}^+$:

$$\int_{\lambda\in\mathbb{R}^+}\prod_{i=1}^{t}\left(\exp\left\{\lambda X_i-\log(1-\mu+\mu e^{\lambda})\right\}\right)dF(\lambda). \tag{51}$$

In particular, Howard et al. (2021) take $F(\lambda)$ to be a beta distribution so that the integral (51) can be computed in closed-form. Using (51) in Step (b) in Theorem 1 yields the "Bernoulli [HRMS20]" confidence sequence.

There are yet other improvements of Hoeffding's inequality, for example one that goes by the name of Kearns-Saul (Kearns and Saul, 1998) but was incidentally noted in Hoeffding's original paper itself. This inequality, and other variants, are looser than the sub-Bernoulli bound and so we exclude them here; see Howard et al. (2020) for more details. Most importantly, none of these adapt to the true underlying variance of the random variables, unlike most of our new techniques.

A-Bentkus [KZ21] We also compared our bounds against the "adaptive Bentkus confidence sequence" (A-Bentkus) due to Kuchibhotla and Zheng (2021, Section 3.5). These combine a maximal version of Bentkus et al.'s concentration inequality (Kuchibhotla and Zheng, 2021, Theorem 1) with the "stitching" technique Zhao et al. (2016); Mnih et al. (2008); Howard et al. (2021) — a method to obtain infinite-horizon concentration inequalities by taking a union bound over exponentially-spaced *finite* time horizons.

D.2. Fixed-time confidence intervals (with replacement)

For the fixed-time CIs included from this paper, we have

- **PrPl-EB-CI**: Predictable plug-in empirical Bernstein CI (Remark 1); and
- **Hedged-CI**: Hedged capital process CI (Remark 3).

These were compared against CIs due to Hoeffding (1963), Maurer and Pontil (2009), Anderson (1969), and Bentkus (2004) which we now recall.

H-CI [H63] These intervals refer to the CIs based on Hoeffding's classical concentration inequalities (Hoeffding, 1963). Specifically, for a sample size $n \geq 1$, "H-CI [H63]" refers to the CI,

$$\frac{1}{n}\sum_{i=1}^{n}X_i \pm \sqrt{\frac{\log(2/\alpha)}{2n}}.$$

Anderson [A69] These intervals refer to the confidence intervals due to Anderson (1969) which take a unique approach by considering the entire sample cumulative distribution function, rather than just the mean and variance. Consequently, however, Anderson's CIs require iid observations, rather than the more general setup we

consider. We nevertheless find that even in the iid setting, our approach outperforms Anderson's.

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} P$ are $[0,1]$-bounded with mean $\mathbb{E}_P(X_1) = \mu$. Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics of X_1^n with the convention that $X_{(0)} := 0$ and $X_{(n+1)} := 1$. Following the notation of Learned-Miller and Thomas (2019), Anderson's CI is given by

$$\left[\sum_{i=1}^{n} u_i^{\mathrm{DKW}} \left(-X_{(n-(i+1))} + X_{(n-i)}\right),\ 1 - \sum_{i=1}^{n} u_i^{\mathrm{DKW}} \left(X_{(i+1)} - X_{(i)}\right)\right],$$

where $u_i^{\mathrm{DKW}} = \left(i/n - \sqrt{\log(2/\alpha)/2n}\right) \vee 0$. Learned-Miller and Thomas (2019, Theorem 2) show that Anderson's CI is always tighter than Hoeffding's. The authors also introduce a bound which is strictly tighter than Anderson's which they conjecture has valid $(1-\alpha)$-coverage, but we do not compare to this bound here.

EB-CI [MP09] The empirical Bernstein CI of Maurer and Pontil (2009) is given by

$$\frac{1}{n}\sum_{i=1}^{n} X_i \pm \sqrt{\frac{2\widehat{\sigma}^2 \log(4/\alpha)}{n}} + \frac{7\log(4/\alpha)}{3(n-1)},$$

and $\widehat{\sigma}^2$ is the sample variance.

Bentkus-CI [B04] Bentkus' confidence interval requires an a-priori upper bound on $\mathrm{Var}(X_i)$ for each i. As alluded to in the introduction, we do not consider concentration bounds which require knowledge of the variance. However, since we assume $X_i \in [0,1]$, we have the trivial upper bound, $\mathrm{Var}(X_i) \leq \frac{1}{4}$, which we implicitly use throughout our computation of Bentkus' confidence interval.

Define the independent, mean-zero random variables $(G_i)_{i=1}^n$ as

$$G_i := \begin{cases} -\frac{1}{4} & \text{w.p. } \frac{4}{5} \\ 1 & \text{w.p. } \frac{1}{5} \end{cases},$$

an important technical device which has appeared in seminal works by Hoeffding (1963, Equation (2.14)) and Bennett (1962, Equation (10)). Then the "Bentkus-CI" is

$$\frac{1}{n}\sum_{i=1}^{n} X_i \pm \frac{W_\alpha^\star}{n},$$

where $W_\alpha^\star \in [0,n]$ is given by the value of W_α such that

$$\inf_{y \in [0,n]\ :\ y \leq W_\alpha} \frac{\mathbb{E}\left[\sum_{i=1}^n (G_i - y)_+^2\right]}{(W_\alpha - y)_+^2} = \alpha.$$

Efficient algorithms have been developed to solve the above (Bentkus et al., 2006, Section 9), (Kuchibhotla and Zheng, 2021).

PTL-ℓ_2 [PTL21] The work by Phan et al. (2021) proposes an interesting but computationally intensive approach to constructing confidence intervals for means of iid bounded random variables. Specifically, we will focus on their tightest bound (according to (Phan et al., 2021, Figure 4)) which makes use of the ℓ_2 norm in its derivation (and which we thus refer to as PTL-ℓ_2).

For example, computing PTL-ℓ_2 confidence intervals¶ from a sample $X_1, \ldots, X_{300} \sim \text{Unif}[0,1]$ of $n = 300$ uniformly distributed random variables took upwards of 11 minutes while our betting confidence interval (Remark 3) took less than 0.5 seconds. For this reason, we conduct a small-scale simulation of sample sizes 5-200 (see Figure 16). We find that PTL-ℓ_2 performs extremely well for the low-variance continuous distribution Beta(10, 30) but poorly for sample sizes closer to 200 for Bernoulli data. Nevertheless, PTL-ℓ_2 requires i.i.d. data (while we only require boundedness and conditional mean μ) and PTL-ℓ_2 does not have time-uniform or without-replacement analogues.

D.3. Time-uniform confidence sequences (without replacement)

The WoR CSs which were introduced in this paper include

- **Hedged-WoR**: Without replacement hedged capital process (Theorem 4); and
- **hgKelly-WoR**: Without replacement analogue of hgKelly (Proposition 5).

The CSs labeled "H-WoR [WR20]" and "EB-WoR [WR20]" are the without-replacement Hoeffding- and empirical Bernstein-type CSs due to Waudby-Smith and Ramdas (2020) which we recall now.

H-WoR [WR20] Define the weighted WoR mean estimator and the Hoeffding-type λ-sequence,

$$\widehat{\mu}_t^{\text{WoR}}(\lambda_1^t) := \frac{\sum_{i=1}^t \lambda_i (X_i + \frac{1}{N-i+1}\sum_{j=1}^{i-1} X_j)}{\sum_{i=1}^t \lambda_i (1 + \frac{i-1}{N-i+1})}, \quad \text{and} \quad \lambda_t := \sqrt{\frac{8\log(2/\alpha)}{t\log(t+1)}} \wedge 1,$$

respectively. Then "H-CS [WR20]" refers to the WoR Hoeffding-type CS,

$$\widehat{\mu}_t^{\text{WoR}}(\lambda_1^t) \pm \frac{\sum_{i=1}^t \psi_H(\lambda_i) + \log(2/\alpha)}{\sum_{i=1}^t \lambda_i \left(1 + \frac{i-1}{N-i+1}\right)}.$$

EB-WoR [WR20] Analogously to the Hoeffding-type CSs, "EB-CS [WR20]" corresponds to the empirical Bernstein-type CSs for sampling WoR due to Waudby-Smith and Ramdas (2020). These CSs take the form

$$\widehat{\mu}_t^{\text{WoR}}(\lambda_1^t) \pm \frac{\sum_{i=1}^t 4(X_i - \widehat{\mu}_{i-1})^2 \psi_E(\lambda_i) + \log(2/\alpha)}{\sum_{i=1}^t \lambda_i \left(1 + \frac{i-1}{N-i+1}\right)},$$

¶We used code by Phan et al. (2021) with their default tuning parameters, available at github.com/myphan9/small_sample_mean_bounds.

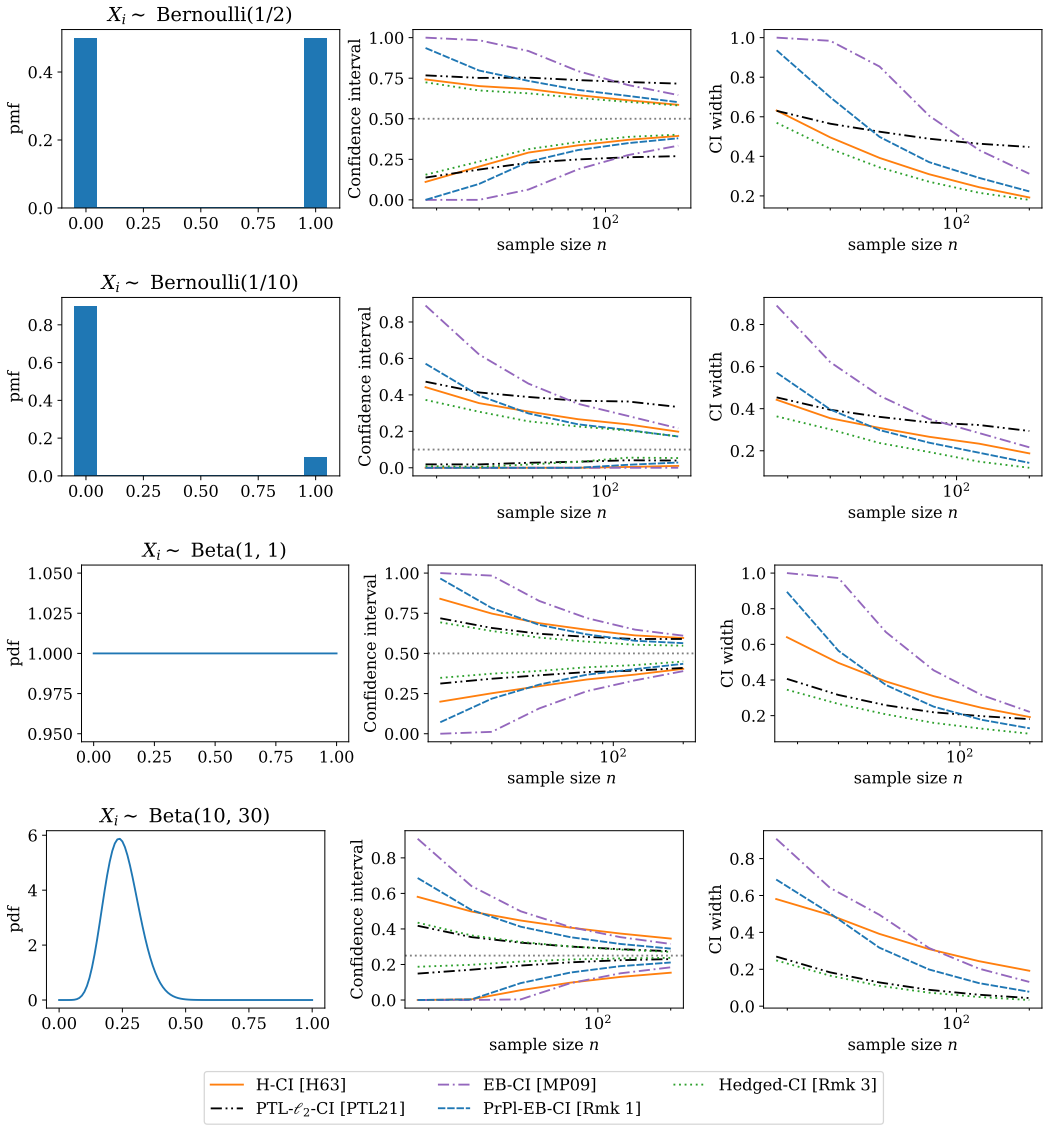


Figure 16. Various with-replacement fixed-time confidence intervals, including that of Phan et al. (2021) (PTL-ℓ_2-CI). While PTL-ℓ_2-CI performs very well in the Beta(10, 30) regime, it appears to suffer for Bernoulli(1/2) with larger n. In any case, PTL-ℓ_2-CI relies on iid data, while the other four methods do not.

where in this case, we have

$$\lambda_t := \sqrt{\frac{2\log(2/\alpha)}{\widehat{\sigma}_{t-1}^2 t\log(t+1)}} \wedge \frac{1}{2}, \quad \widehat{\sigma}_t^2 := \frac{1/4 + \sum_{i=1}^t (X_i - \widehat{\mu}_i)^2}{t+1}, \quad \text{and} \quad \widehat{\mu}_t := \frac{1}{t}\sum_{i=1}^t X_i. \tag{52}$$

D.4. Fixed-time confidence intervals (without replacement)

The only fixed-time CI introduced in this paper is **Hedged-WoR-CI**: the without-replacement hedged capital process CI described in Section 5. The other two are both due to Waudby-Smith and Ramdas (2020) which we describe now.

H-WoR-CI [WR20] This corresponds to the CI described in Corollary 3.1 of Waudby-Smith and Ramdas (2020). This has the form

$$\widehat{\mu}_n^{\mathrm{WoR}} \pm \frac{\sqrt{\frac{1}{2}\log(2/\alpha)}}{\sqrt{n} + \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{i-1}{N-i+1}}.$$

EB-WoR-CI [WR20] Similarly, this CI corresponds to that described in Corollary 3.2 of Waudby-Smith and Ramdas (2020). Specifically, "EB-WoR-CI [WR20]" is defined as

$$\widehat{\mu}_n^{\mathrm{WoR}}(\lambda_1^n) \pm \frac{\sum_{i=1}^n 4(X_i - \widehat{\mu}_{i-1})^2 \psi_E(\lambda_i) + \log(2/\alpha)}{\sum_{i=1}^n \lambda_i \left(1 + \frac{i-1}{N-i+1}\right)},$$

where

$$\lambda_t := \sqrt{\frac{2\log(2/\alpha)}{n\widehat{\sigma}_{t-1}^2}} \wedge \frac{1}{2}, \quad \widehat{\sigma}_t^2 := \frac{1/4 + \sum_{i=1}^t (X_i - \widehat{\mu}_i)^2}{t+1}, \text{ and } \widehat{\mu}_t := \frac{\frac{1}{2} + \sum_{i=1}^t X_i}{t+1}, \tag{53}$$

and $\widehat{\mu}_n^{\mathrm{WoR}}$ is defined as

$$\widehat{\mu}_t^{\mathrm{WoR}}(\lambda_1^t) := \frac{\sum_{i=1}^t \lambda_i (X_i + \frac{1}{N-i+1}\sum_{j=1}^{i-1} X_j)}{\sum_{i=1}^t \lambda_i (1 + \frac{i-1}{N-i+1})}.$$

D.5. Betting "confidence distributions": confidence sets at several resolutions

Figures 17 and 18 demonstrate two tools to visualize CSs at various α and t.

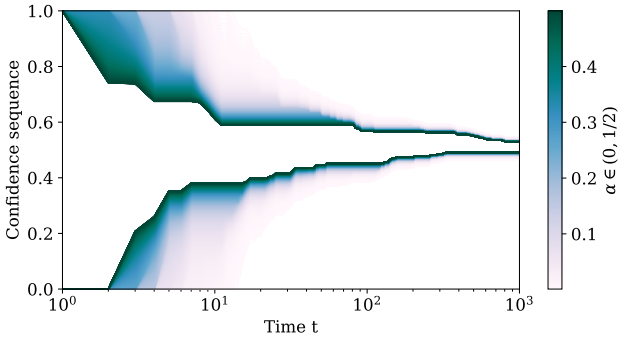


Figure 17. This plot shows the aGRAPA CS for all $\alpha \in [0, 1/2]$ under Unif$[0, 1]$ data.

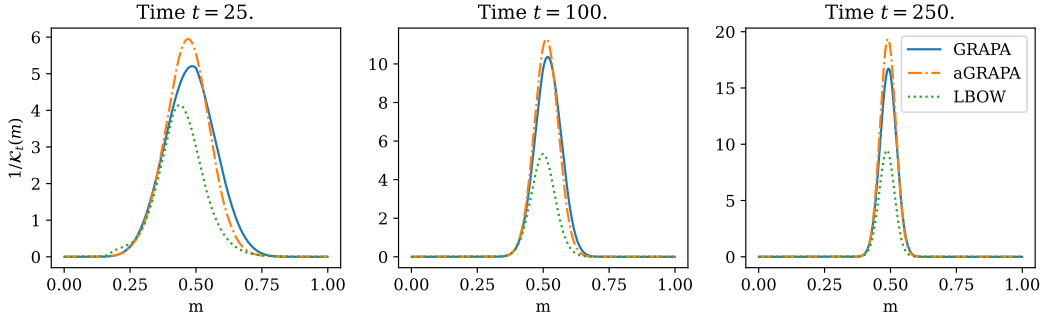


Figure 18. Here we plot the inverse wealth $1/\mathcal{K}_t(m)$ in game m against $m \in [0, 1]$, at $t = 25, 100, 250$ for three different betting strategies. Note the different y-axis scales. Despite not being normalized to yield a "confidence distribution", this is a useful visual tool. For example, the mode in each plot signifies the m against which we have minimum wealth, which is a reasonable point estimator for μ. Further, the superlevel set for any $\alpha \in [0, 1]$ yields exactly the $(1 - \alpha)$-CS for μ (for that corresponding time and strategy) since it yields all m with wealth less than $1/\alpha$. Last, for any $m \in [0, 1]$, the height (truncated at one) is anytime-valid p-value for the null hypothesis that the mean equals m.

E. Additional theoretical results

E.1. Betting confidence sets are tighter than Hoeffding

In this section, we demonstrate that the betting approach can dominate Hoeffding for sufficiently large sample sizes. First, we show that for any $x, m \in (0, 1)$ and any $\lambda \in \mathbb{R}$, then $\gamma \equiv \gamma^m(\lambda)$ can be set as

$$\gamma^m(\lambda) := \exp\left\{-m\lambda - \lambda^2/8\right\}(\exp(\lambda) - 1),$$

so that

$$H^m(x) := \underbrace{\exp\left\{\lambda(x - m) - \lambda^2/8\right\}}_{\text{Hoeffding term}} \leq \underbrace{1 + \gamma(x - m)}_{\text{Capital process term}} =: \mathcal{K}^m(x)$$

for any $x, m \in [0,1]$. In particular, the Hoeffding-type and capital process supermartingales are built from precisely the above terms, respectively, and so if $H^m(x) \leq \mathcal{K}^m(x)$ for any $x \in [0,1]$, then their respective supermartingales will satisfy the same inequality almost surely.

PROPOSITION 6 (CAPITAL PROCESS DOMINATES HOEFFDING PROCESS). *Suppose $x, m \in [0,1]$ and $\lambda \in \mathbb{R}$. Then there exists $\gamma^m(\lambda) \in \mathbb{R}$ such that*

$$H^m(x) := \exp\left(\lambda(x-m) - \lambda^2/8\right) \leq 1 + \gamma^m(\lambda)(x-m) =: \mathcal{K}^m(x).$$

Note that Proposition 6 alone does not confirm that the Hoeffding-based CIs will be dominated by capital process-based CIs since γ must be within $[-1/(1-m), 1/m]$ for $\mathcal{K}^m(x)$ to be nonnegative. However, it is easy to verify that for all $\lambda \in [-0.45, 0.45]$, we have that $\gamma \in [-1,1]$ and thus $\mathcal{K}^m(x) \geq 0$. When constructing a Hoeffding-type $(1-\alpha)$-confidence interval, for example, one would set $\lambda_n^H := \sqrt{8\log(2/\alpha)/n}$, making $\lambda_n^H \in [-0.45, 0.45]$ whenever $n \geq 40\log(2/\alpha)$, in which case a capital process-based CI will dominate a Hoeffding-based CI almost surely.

PROOF (PROPOSITION 6). We prove the result for $\lambda \geq 0$ and remark that this implies the result for the case when $\lambda \leq 0$ by considering $(1-x)$ and $(1-m)$ instead of x and m, respectively.

The proof proceeds in 3 steps. First, we consider the line segment $L^m(x)$ connecting $H^m(0)$ and $H^m(1)$ and note that by convexity of $H^m(x)$, we have that $H^m(x) \leq L^m(x)$ for all $x \in [0,1]$. We then find the slope of this line segment and set γ to this value so that the line $\mathcal{K}^m(x) := 1 + \gamma(x-m)$ has the same slope as $L^m(x)$. Finally, we demonstrate that $L^m(0) \leq \mathcal{K}^m(0)$, and conclude that $H^m(x) \leq L^m(x) \leq \mathcal{K}^m(x)$ for all $x \in [0,1]$.

Step 1. Note that $H^m(x)$ is a convex function in $x \in [0,1]$, and thus

$$\forall x \in [0,1],\ H^m(x) \leq H^m(0) + [H^m(1) - H^m(0)]\, x =: L^m(x).$$

Step 2. Observe that the slope of $L^m(x)$ is $H^m(1) - H^m(0)$. Setting $\gamma := H^m(1) - H^m(0)$ we have that $\mathcal{K}^m(x)$ and $L^m(x)$ are parallel.

Step 3. It remains to show that $\mathcal{K}^m(0) \geq L^m(0) \equiv H^m(0)$ for every $m \in [0,1]$. Consider the following equivalent statements:

$$\begin{aligned}
&\mathcal{K}^m(0) \geq H^m(0) \\
\iff &1 - m\,[H^m(1) - H^m(0)] \geq H^m(0) \\
\iff &1 - m\exp\left(\lambda - \lambda m - \lambda^2/8\right) \geq (1-m)\exp\left(-\lambda m - \lambda^2/8\right) \\
\iff &1 \geq \exp\left(-\lambda m - \lambda^2/8\right)[1 - m + m\exp(\lambda)] \\
\iff &\exp\left(\lambda m + \lambda^2/8\right) \geq [1 - m + m\exp(\lambda)] \\
\iff &a(\lambda) := \exp\left(\lambda m + \lambda^2/8\right) - [1 - m + m\exp(\lambda)] \geq 0.
\end{aligned}$$

Now, note that a is smooth and $a(0) = 0$ and so it suffices to show that its derivative $a'(\lambda) \geq 0$ for all $\lambda \geq 0$. To this end, consider the following equivalent statements.

$$\begin{aligned} & a'(\lambda) \equiv \left(m + \frac{\lambda}{4}\right) \exp\left(\lambda m + \lambda^2/8\right) - m \exp(\lambda) \geq 0 \\ \iff & \left(m + \frac{\lambda}{4}\right) \exp\left(\lambda m + \lambda^2/8\right) \geq m \exp(\lambda) \\ \iff & \ln\left(1 + \frac{\lambda}{4m}\right) + \lambda m + \lambda^2/8 \geq \lambda \\ \iff & b(\lambda) := \ln\left(1 + \frac{\lambda}{4m}\right) + \lambda m + \lambda^2/8 - \lambda \geq 0, \end{aligned}$$

and hence it suffices to show that $b(\lambda) \geq 0$. Similar to $a(\lambda)$, we have that $b(0) = 0$ and so it suffices to show that its derivative, $b'(\lambda) \geq 0$ for all $\lambda \geq 0$. Indeed,

$$\begin{aligned} & b'(\lambda) \equiv \frac{1}{4m + \lambda} + m + \frac{\lambda}{4} - 1 \geq 0 \\ \iff & c(\lambda) := 1 + m(4m + \lambda) + \frac{\lambda}{4}(4m + \lambda) - 4m - \lambda \geq 0 \end{aligned}$$

Since $c(\lambda)$ is a convex quadratic, it is straightforward to check that

$$\operatorname*{argmin}_{\lambda \in \mathbb{R}} c(\lambda) = 2 - 4m,$$

and that $c(2 - 4m) = 0$. In conclusion, if we set $\gamma \equiv \gamma^m(\lambda)$ as

$$\gamma^m(\lambda) := H^m(1) - H^m(0) = \exp\left\{-m\lambda - \lambda^2/8\right\}(\exp(\lambda) - 1),$$

then $H^m(x) \leq \mathcal{K}^m(x) := 1 + \gamma^m(\lambda)(x - m)$ for every $m \in [0, 1]$. This completes the proof. □

E.2. Optimal convergence of betting confidence sets

In Section B, it was mentioned that for nonnegative martingales, Ville's inequality is nearly an equality and hence martingale-based CSs are nearly tight in a time-uniform sense. However, it is natural to wonder what other theoretical guarantees betting CSs/CIs can have in addition to their empirical performance. In the time-uniform setting, CSs for the mean cannot attain widths which scale faster than $\asymp \sqrt{\log\log t/t}$, due to the law of the iterated logarithm. Similarly, fixed-time CIs cannot scale faster than $\asymp 1/\sqrt{n}$. In this section, we show that it is possible to choose betting strategies such that the resulting CSs and CIs scale at the optimal rates of $O(\sqrt{\log\log t/t})$ and $O(1/\sqrt{n})$, respectively.

E.2.1. An iterated logarithm betting confidence sequence

We will establish the law of the iterated logarithm (LIL) convergence rate by carefully constructing a capital process martingale whose resulting CS is — for sufficiently large t — tighter than a larger CS which itself attains the required LIL rate.

Before stating the result in Proposition 7, let $\zeta(s) := \sum_{k=1}^{\infty} \frac{1}{k^s}$ be the Riemann zeta function and for each $k \in \{1, 2, \dots\}$, define

$$\lambda_k := \sqrt{\frac{8 \log\left(k^s \zeta(s)\right)}{\eta^{k+1/2}}}, \quad \text{and}$$
$$\gamma_k(m) = \exp\left\{-m\lambda_k - \lambda_k^2/8\right\} (\exp(\lambda_k) - 1) \wedge 1,$$

where $\eta > 1$ is some user-chosen constant. Let k_t denote the (unique) integer such that $\log_\eta t \le k_t \le \log_\eta t + 1$. Define the process

$$\mathcal{K}_t^{\mathcal{L}} := \frac{1}{2}\mathcal{K}_t^{\mathcal{L}+}(m) + \frac{1}{2}\mathcal{K}_t^{\mathcal{L}-}(m)$$
$$\text{where} \quad \mathcal{K}_t^{\mathcal{L}+}(m) := \frac{1}{k_t^s \zeta(s)} \prod_{i=1}^{t} (1 + \gamma_{k_t}(X_i - m)) \quad \text{and}$$
$$\mathcal{K}_t^{\mathcal{L}-}(m) := \frac{1}{k_t^s \zeta(s)} \prod_{i=1}^{t} (1 - \gamma_{k_t}(X_i - m)).$$

Note that $\mathcal{K}_t^{\mathcal{L}+}(m)$ and $\mathcal{K}_t^{\mathcal{L}-}(m)$ are both upper-bounded by the infinite mixtures

$$\mathcal{K}_t^{\mathcal{L}+}(m) \le \sum_{k=1}^{\infty} \frac{1}{k^s \zeta(s)} \prod_{i=1}^{t} (1 + \gamma_k(X_i - m)) \quad \text{and} \tag{54}$$
$$\mathcal{K}_t^{\mathcal{L}-}(m) \le \sum_{k=1}^{\infty} \frac{1}{k^s \zeta(s)} \prod_{i=1}^{t} (1 - \gamma_k(X_i - m)), \tag{55}$$

which themselves form nonnegative martingales when $m = \mu$ by Fubini's theorem. Consequently,

$$C_t^{\mathcal{L}} := \left\{ m \in [0,1] : \mathcal{K}_t^{\mathcal{L}}(m) < \frac{1}{\alpha} \right\}$$

forms a $(1-\alpha)$-CS for μ. The following proposition establishes the LIL rate of $C_t^{\mathcal{L}}$.

PROPOSITION 7. *The CS $(C_t^{\mathcal{L}})_{t=1}^{\infty}$ has a width of $O(\sqrt{\log\log t/t})$, meaning*

$$\nu(C_t^{\mathcal{L}}) = O\left(\sqrt{\frac{\log\log t}{t}}\right),$$

where ν is the Lebesgue measure.

PROOF. The proof proceeds in three steps. In Step 1, we construct a distinct but related CS (which we will denote by $(C_t^{\times})_{t=1}^{\infty}$) via the stitching technique (Howard et al., 2021). In Step 2, we demonstrate that this stitched CS achieves the desired rate by deriving an analytically tractible superset whose width scales as $O(\sqrt{\log\log t/t})$. Finally, in Step 3, we will show that the stitched CS $C_t^{\times}$ is a superset of $C_t^{\mathcal{L}}$ for all t sufficiently large, thus implying the final result.

Step 1. Constructing the stitched CS $C_t^{\times}$: In the language of betting, the idea behind stitching is to first divide one's capital up into infinitely many portions $w_1, w_2, \dots$ such that $\sum_{k=1}^{\infty} w_k = 1$, and then place a constant bet λ_k using a capital of w_k on a designated epoch of time, which will be chosen to be geometrically spaced. In what follows, the portions w_k will be given by $w_k = \frac{1}{\zeta(s)k^s}$, and we will divide time $\{1, 2, 3, \dots\}$ up into epochs demarcated by the endpoints η^{k-1} and η^k for each $k \in \{1, 2, 3, \dots\}$ and for some user-specified $\eta > 1$ (e.g. $\eta = 1.1$). The constant bets λ_k will be chosen so that they are effective between η^{k-1} and η^k and lead to $O(\sqrt{\log\log t/t})$ widths after being combined across epochs.

The construction of the stitched boundary essentially follows (a simplified version of) the proof of Theorem 1 in Howard et al. (2021, Section A.1), but we present the derivation here for completeness. Consider the Hoeffding-type process for a fixed $\lambda \in \mathbb{R}$:

$$M_t^{\lambda}(m) := \exp\left\{\lambda S_t(m) - t\lambda^2/8\right\}, \tag{56}$$

where $S_t(m) := \sum_{i=1}^{t}(X_i - m)$. As discussed in Section 3, $M_t(\mu)$ forms a test supermartingale, and hence by Ville's inequality we have

$$P\left(\exists t \geq 1 : S_t(\mu) \geq \underbrace{\frac{r + t\lambda^2/8}{\lambda}}_{g_{\lambda,r}(t)}\right) \leq e^{-r}.$$

We have typically used $r = \log(1/\alpha)$ throughout the paper, but the above alternative notation will help in the following discussion. Using the notation of Howard et al. (2021, Section A.1), define the boundary above as $g_{\lambda,r}(t) := (r + t\lambda^2/8)/\lambda$, and let

$$\lambda_k := \sqrt{\frac{8r_k}{\eta^{k-1/2}}},$$
$$\text{where} \quad r_k := \log\left(\frac{k^s\zeta(s)}{\alpha/2}\right).$$

Some algebra will reveal that plugging the above choices of λ_k and r_k into $g_{\lambda,r}(t)$ yields

$$g_{\lambda_k,r_k}(t) := \sqrt{\frac{r_k t}{8}}\left(\sqrt{\frac{\eta^{k-1/2}}{t}} + \sqrt{\frac{t}{\eta^{k-1/2}}}\right),$$

resulting in the following concentration inequality for each k:

$$P\left(\exists t \geq 1 : S_t(\mu) \geq g_{\lambda_k,r_k}(t)\right) \leq \exp\{-r_k\}.$$

Let k_t denote the (unique) epoch number such that $\eta^{k_t-1} \leq t \leq \eta^{k_t}$ (i.e. such that $\log_\eta t \leq k_t \leq \log_\eta t + 1$). Now, we take a union bound over $k = 1, 2, 3, \dots$ resulting in the following boundary,

$$P\left(\exists t \geq 1 : S_t(\mu) \geq g_{\lambda_{k_t},r_{k_t}}(t)\right) \leq \sum_{k=1}^{\infty}\exp\{-r_k\} = \frac{\alpha/2}{\zeta(s)}\underbrace{\sum_{k=1}^{\infty}\frac{1}{k^s}}_{\zeta(s)} = \alpha/2.$$

Repeating all of the previous steps for $-S(\mu)$ and taking a union bound, we arrive at the $(1-\alpha)$ stitched CS $(C_t^{\times})_{t=1}^{\infty}$ given by

$$C_t^{\times} := \left(\frac{1}{t} \sum_{i=1}^{t} X_i \pm \frac{g_{\lambda_{k_t}, r_{k_t}}(t)}{t} \right),$$

with the guarantee that $P(\exists t \geq 1 : \mu \notin C_t^{\times}) \leq \alpha$.

Step 2. Demonstrating that $C_t^{\times}$ achieves the desired LIL width: Now, we will simply upper-bound $g_{\lambda_{k_t}, r_{k_t}}(t)$ by an analytical boundary depending explicitly on t (rather than implicitly through k_t) to see that it achieves the desired LIL width. First, notice that $\sqrt{\eta^{k_t-1/2}/t} + \sqrt{t/\eta^{k_t-1/2}}$ is uniquely minimized when $t = \eta^{k_t-1/2}$ and hence its maximum on the interval $(\eta^{k_t-1}, \eta^{k_t})$ must be at the endpoints. Therefore, $\sqrt{\eta^{k_t-1/2}/t} + \sqrt{t/\eta^{k_t-1/2}} \leq \eta^{1/4} + \eta^{-1/4}$ and thus for each k, we have

$$g_{\lambda_{k_t}, r_{k_t}}(t) \leq \sqrt{\frac{r_{k_t} t}{8}} \left(\eta^{1/4} + \eta^{-1/4} \right) \quad \text{for all } \eta^{k_t-1} \leq t \leq \eta^{k_t}.$$

Furthermore, for all $\eta^{k_t-1} \leq t \leq \eta^{k_t}$, we have that $k_t \leq \log_\eta t + 1$. Applying this inequality to the above, we obtain the final bound which does not depend on k,

$$g_{\lambda_{k_t}, r_{k_t}}(t) \leq \sqrt{\frac{t \log\left(2\left(\log_\eta t + 1\right)^s \zeta(s)/\alpha\right)}{8}} \left(\eta^{1/4} + \eta^{-1/4} \right) \quad \text{for all } k.$$

In conclusion, we have that

$$C_t^{\times} \subseteq \left(\frac{1}{t} \sum_{i=1}^{t} X_i \pm \sqrt{\frac{\log\left(2\left(\log_\eta t + 1\right)^s \zeta(s)/\alpha\right)}{8t}} \left(\eta^{1/4} + \eta^{-1/4} \right) \right),$$

and thus $C_t^{\times} = O\left(\sqrt{\log\log t / t}\right)$, as desired.

Step 3. Showing that $C_t^{\mathcal{L}} \subseteq C_t^{\times}$ for all t large enough: This step in the proof essentially follows immediately from the discussion in Section E.1. We justified that for $\lambda \geq 0$, setting γ as

$$\gamma = \exp\left\{-m\lambda - \lambda^2/8\right\} \left(\exp(\lambda) - 1\right) \wedge 1,$$

yields $1 + \gamma(x - m) \geq \exp\left\{\lambda(x-m) - \lambda^2/8\right\}$ for all $x, m \in [0,1]$ if λ is sufficiently small (i.e. so that γ is not relying on truncation at 1). Since λ_k is decreasing in t, it follows that for t sufficiently large,

$$\prod_{i=1}^{t} (1 + \gamma_{k_t}(X_i - m)) \geq \exp\left\{\lambda_{k_t} S_t(m) - \lambda_{k_t}^2/8\right\} \quad \text{almost surely.}$$

Therefore, for t sufficiently large,

$$\begin{aligned}\mathcal{K}_t^{\mathcal{L}+}(m) &:= \frac{1}{k_t^s \zeta(s)} \prod_{i=1}^{t} (1 + \gamma_{k_t}(X_i - m)) \\ &\geq \frac{1}{k_t^s \zeta(s)} \exp\left\{\lambda_{k_t} S_t(m) - \lambda_{k_t}^2/8\right\} =: H_t^{\infty+}(m)\end{aligned}$$

and similarly for $K_t^{\mathcal{L}-}(m)$,

$$\mathcal{K}_t^{\mathcal{L}-}(m) \geq \frac{1}{k_t^s \zeta(s)} \exp\left\{-\lambda_{k_t} S_t(m) - \lambda_{k_t}^2/8\right\} =: H_t^{\infty-}(m).$$

Therefore, for sufficiently large t, we have

$$\begin{aligned}C_t^{\mathcal{L}} &:= \left\{m \in [0,1] : \mathcal{K}_t^{\mathcal{L}}(m) < \frac{1}{\alpha}\right\} \\ &\subseteq \underbrace{\left\{m \in \mathbb{R} : \max\left\{\frac{1}{2} H_t^{\infty+}(m),\ \frac{1}{2} H_t^{\infty-}(m)\right\} < \frac{1}{\alpha}\right\}}_{(\star)}\end{aligned}$$

and it is straightforward to verify that $(\star)$ is precisely $C_t^{\times}$.

In summary, we constructed a CS $C_t^{\times}$ using the stitching technique in Step 1, and then showed that $\nu(C_t^{\times}) = O(\sqrt{\log\log t/t})$ in Step 2. Finally in Step 3, we showed that our discrete mixture betting CS $C_t^{\mathcal{L}}$ is a subset of $C_t^{\times}$ for t sufficiently large, and hence by subadditivity of measures,

$$\nu(C_t^{\mathcal{L}}) = O\left(\sqrt{\frac{\log\log t}{t}}\right),$$

which completes the proof. □

REMARK 6. *Notice that $\mathcal{K}_t^{\mathcal{L}+}$ and $\mathcal{K}_t^{\mathcal{L}-}$ can be made strictly more powerful if they are replaced by adding additional terms, as long as the final sums are upper-bounded by* (54) *and* (55)*, respectively. In particular, any finite sum analogue of* (54) *and* (55) *would have sufficed, as long as $\mathcal{K}_t^{\mathcal{L}+}$ and $\mathcal{K}_t^{\mathcal{L}-}$ form a term in each sum, respectively. We presented $\mathcal{K}_t^{\mathcal{L}+}$ and $\mathcal{K}_t^{\mathcal{L}-}$ in their current forms for the sake of notational (and computational) simplicity.*

E.2.2. The $\sqrt{n}$-convergence of betting CIs

PROPOSITION 8. *Suppose $X_1^n \sim P$ are independent observations from a distribution $P \in \mathcal{P}^{\mu}$ with mean $\mu \in [0,1]$. Let $\lambda_n \in (0,1)$ such that $\lambda_n \asymp 1/\sqrt{n}$. Then the confidence interval,*

$$C_n := \left\{m \in [0,1] : \mathcal{K}_n^{\pm} < \frac{1}{\alpha}\right\} \quad \textit{has an asymptotic width of } O(1/\sqrt{n}).$$

PROOF. Writing out the capital process with positive bets, we have by Lemma 3 that for any $m \in [0,1]$,

$$\begin{aligned}\mathcal{K}_n^+(m) &:= \prod_{i=1}^n (1 + \lambda_n(X_i - m)) \\ &\geq \exp\left(\lambda_n \sum_{i=1}^n (X_i - m) - \psi_E(\lambda_n) \sum_{i=1}^n 4(X_i - m)^2\right) \\ &\geq \exp\left(\lambda_n \sum_{i=1}^n (X_i - m) - 4n\psi_E(\lambda_n)\right) =: B_t^+(m),\end{aligned}$$

and similarly for negative bets,

$$\begin{aligned}\mathcal{K}_n^-(m) &:= \prod_{i=1}^n (1 - \lambda_n(X_i - m)) \\ &\geq \exp\left(-\lambda_n \sum_{i=1}^t (X_i - m) - 4n\psi_E(\lambda_n)\right) =: B_t^-(m).\end{aligned}$$

For any $\theta \in (0,1)$, consider the set,

$$\mathcal{S}_n := \left\{m : B_t^+(m) < \frac{1}{\theta\alpha}\right\} \bigcap \left\{m : B_t^-(m) < \frac{1}{(1-\theta)\alpha}\right\}$$

Now notice that the $1/\alpha$-level set of $\mathcal{K}_n^{\pm}(m) := \max\{\theta\mathcal{K}_n^+(m), (1-\theta)\mathcal{K}_n^-(m)\}$ is a subset of $\mathcal{S}_n$:

$$C_n = \left\{m : \mathcal{K}_n^+(m) < \frac{1}{\theta\alpha}\right\} \bigcap \left\{m : \mathcal{K}_n^-(m) < \frac{1}{(1-\theta)\alpha}\right\} \subseteq \mathcal{S}_n.$$

On the other hand, it is straightforward to derive a closed-form expression for $\mathcal{S}_n$:

$$\left(\frac{\sum_{i=1}^n X_i}{n} - \frac{\log\left(\frac{1}{\theta\alpha}\right) + 4n\psi_E(\lambda_n)}{n\lambda_n}, \frac{\sum_{i=1}^n X_i}{n} + \frac{\log\left(\frac{1}{(1-\theta)\alpha}\right) + 4n\psi_E(\lambda_n)}{n\lambda_n}\right),$$

which in the typical case of $\theta = 1/2$ has the cleaner expression,

$$\frac{\sum_{i=1}^n X_i}{n} \pm \frac{\log(2/\alpha) + 4n\psi_E(\lambda_n)}{n\lambda_n}.$$

As discussed in Section B, we have by two applications of L'Hôpital's rule that $\frac{\psi_E(\lambda_n)}{\psi_H(\lambda_n)} \xrightarrow{n\to\infty} 1$, where $\psi_H(\lambda_n) := \lambda_n^2/8 \asymp 1/n$ and thus the width W_n of $\mathcal{S}_n$ scales as

$$W_n := 2 \cdot \frac{\log(1/\alpha) + 4n\psi_E(\lambda_n)}{n\lambda_n} \asymp \frac{\log(1/\alpha)}{\sqrt{n}} + \frac{4n/n}{\sqrt{n}} \asymp \frac{1}{\sqrt{n}}.$$

Since $C_n \subseteq \mathcal{S}_n$, we have that C_n has a width of $O(1/\sqrt{n})$, which completes the proof. □

Despite these results, the hedged capital CI presented and recommended in Section 4.4 does not satisfy the assumptions of the above proof. In particular, we recommended using the variance-adaptive predictable plug-in,

$$\lambda_t^{\text{PrPl-EB}(n)} := \sqrt{\frac{2\log(2/\alpha)}{n\widehat{\sigma}_{t-1}^2}}, \quad \widehat{\sigma}_t^2 := \frac{1/4 + \sum_{i=1}^t (X_i - \widehat{\mu}_i)^2}{t+1}, \quad \text{and} \quad \widehat{\mu}_t := \frac{1/2 + \sum_{i=1}^t X_i}{t+1}, \tag{57}$$

using a truncation which depends on m,

$$\lambda_t^+(m) := \lambda_t^{\pm} \wedge \frac{c}{m}, \quad \lambda_t^-(m) := -\left(\lambda_t^{\pm} \wedge \frac{c}{1-m}\right), \tag{58}$$

and finally defining the hedged capital process for each $t \in \{1, \ldots, n\}$:

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

Furthermore, the resulting CI is defined as an intersection,

$$\mathfrak{B}_n := \bigcap_{t=1}^n \left\{m \in [0,1] : \mathcal{K}_t^{\pm}(m) < \frac{1}{\alpha}\right\}. \tag{59}$$

All of these tweaks (i.e. making bets predictable, truncating beyond $(0,1)$, and taking an intersection) do not in any way invalidate the type-I error, but we find (through simulations) that they tighten the CIs, especially in low-variance, asymmetric settings (see Figure 19).

E.3. On the width of empirical Bernstein confidence intervals

Recall the predictable plug-in empirical Bernstein confidence interval:

$$C_n^{\text{PrPl-EB}(n)} := \left(\frac{\sum_{i=1}^n \lambda_i X_i}{\sum_{i=1}^n \lambda_i} \pm \frac{\log(2/\alpha) + \sum_{i=1}^n v_i \psi_E(\lambda_i)}{\sum_{i=1}^n \lambda_i}\right),$$

where

$$\lambda_t := \sqrt{\frac{2\log(2/\alpha)}{n\widehat{\sigma}_{t-1}^2}}, \ \widehat{\sigma}_t^2 := \frac{\frac{1}{4} + \sum_{i=1}^t (X_i - \widehat{\mu}_i)^2}{t+1}, \text{and } \widehat{\mu}_t := \frac{\frac{1}{2} + \sum_{i=1}^t X_i}{t+1}.$$

Below, we analyze the asymptotic behavior of the width of $C_n^{\text{PrPl-EB}(n)}$ in the i.i.d. setting. In Proposition 9, we will show that if the data are drawn i.i.d. from a distribution $Q \in \mathcal{Q}^\mu$ having variance σ^2, then the half-width W_n of $C_n^{\text{PrPl-EB}(n)}$ scales as

$$\sqrt{n} W_n \ \equiv \ \sqrt{n}\left(\frac{\log(2/\alpha) + \sum_{i=1}^n v_i \psi_E(\lambda_i)}{\sum_{i=1}^n \lambda_i}\right) \xrightarrow{a.s.} \sigma\sqrt{2\log(2/\alpha)}, \tag{60}$$

and hence the width is asymptotically proportional to the standard deviation.

First, let us prove a few lemmas about *nonrandom* sequences of numbers, which will be helpful in what follows. These are simple facts for which we could not find a proof to reference, so we prove them below for completeness.

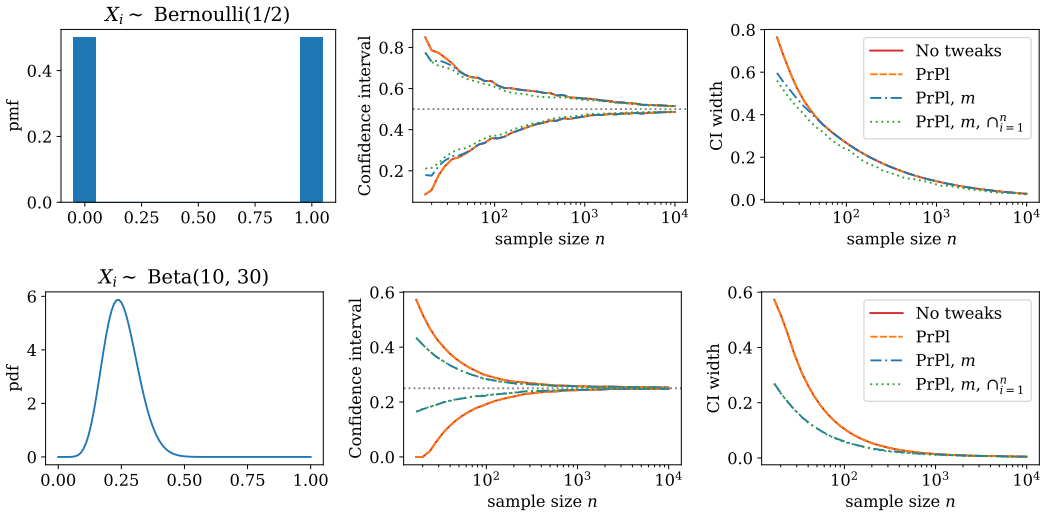


Figure 19. Hedged capital CIs with various added tweaks. The CIs labeled "No tweaks" refer to those which satisfy the conditions of Proposition 8. The other three plots differ in which "tweaks" have been added. Those with "PrPl" in the legend use the predictable plug-in approach defined in (57); those with m in the legend have been truncated using m as outlined in (58); finally, the plots with $\cap_{i=1}^n$ in their legends had their running intersections taken as in (59).

LEMMA 4. *Suppose $(a_n)_{n=1}^\infty$ is a sequence of real numbers such that $a_n \to a$. Then their cumulative average also converges to a, meaning that $\frac{1}{n}\sum_{i=1}^n a_i \to a$.*

PROOF. Let $\epsilon > 0$ and choose $N \equiv N_\epsilon \in \mathbb{N}$ such that whenever $n \geq N$, we have

$$|a_n - a| < \epsilon. \tag{61}$$

Moreover, choose

$$M \equiv M_N > \frac{\sum_{i=1}^N |a_i - a|}{\epsilon} \tag{62}$$

and note that

$$\frac{n - N - 1}{n} < 1. \tag{63}$$

Let $n \geq \max\{N, M\}$. Then we have by the triangle inequality,

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^n (a_i - a)\right| &\leq \frac{1}{n}\sum_{i=1}^N |a_i - a| + \frac{1}{n}\sum_{i=N+1}^n |a_i - a| \\
&\leq \frac{1}{n}\sum_{i=1}^N |a_i - a| + \frac{1}{n}(n - N - 1)\epsilon && \text{by (61)} \\
&\leq 2\epsilon && \text{by (62) and (63)},
\end{aligned}$$

which can be made arbitrarily small. This completes the proof of Lemma 4. □

LEMMA 5. *Let $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=1}^{\infty}$ be sequences of numbers such that*

$$a_n \to 0 \quad \textit{and} \tag{64}$$

$$|b_n| \le C \quad \textit{for some } C \ge 0 \textit{ and for all } n \ge 1. \tag{65}$$

Then $a_n b_n \to 0$. Further, if (A_n) is a sequence of random variables such that $A_n \to 0$ almost surely, then $A_n b_n \to 0$ almost surely.

The proof is trivial, since $|A_n b_n| \le C|A_n|$ which converges to zero almost surely. □

Now, we prove that a modified variance estimator is consistent.

LEMMA 6. *Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} Q \in \mathcal{Q}^{\mu}$ with $\mathrm{Var}(X_i) = \sigma^2$. Then the modified variance estimator*

$$\widehat{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu}_{i-1})^2$$

converges to σ^2, Q-almost surely.

PROOF. By direct substitution,

$$\widehat{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu}_{i-1})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu + \mu - \widehat{\mu}_{i-1})^2$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2}_{\xrightarrow{a.s.}\sigma^2} - \underbrace{\frac{2}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu}_{i-1})(\widehat{\mu}_{i-1} - \mu)}_{(\star)} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\mu - \widehat{\mu}_{i-1})^2}_{(\star\star)}.$$

Now, note that $\widehat{\mu}_{i-1} - \mu \xrightarrow{a.s.} 0$ and $|X_i - \widehat{\mu}_{i-1}| \le 1$ for each i. Therefore, by Lemma 5, $(X_i - \widehat{\mu}_{i-1})(\widehat{\mu}_{i-1} - \mu) \xrightarrow{a.s.} 0$, and by Lemma 4, $(\star) \xrightarrow{a.s.} 0$. Furthermore, we have that $(\mu - \widehat{\mu}_{i-1})^2 \xrightarrow{a.s.} 0$ and so by another application of Lemma 4, we have $(\star\star) \xrightarrow{a.s.} 0$. This completes the proof of Lemma 6. □

Next, let us analyze the second term in the numerator in the margin of $C_n^{\text{PrPl-EB}(n)}$,

$$\frac{\log(2/\alpha) + \sum_{i=1}^{n} v_i \psi_E(\lambda_i)}{\sum_{i=1}^{n}\lambda_i}. \tag{66}$$

LEMMA 7. *Under the same assumptions as Lemma 6,*

$$\sum_{i=1}^{n} v_i \psi_E(\lambda_i) \xrightarrow{a.s.} \log(2/\alpha).$$

PROOF. Recall that $\frac{\psi_E(\lambda)}{\psi_H(\lambda)} \xrightarrow{\lambda \to 0} 1$, and $\widehat{\sigma}_t^2 \xrightarrow{t\to\infty} \sigma^2$. By definition of λ_i, we have that $\lambda_i \xrightarrow{a.s.} 0$ and thus we may also write

$$\frac{\psi_E(\lambda_i)}{\psi_H(\lambda_i)} = 1 + R_i \quad \text{and} \tag{67}$$

$$\sqrt{\frac{\sigma^2}{\widehat{\sigma}_t^2}} = 1 + R_i' \tag{68}$$

for some $R_i, R_i' \xrightarrow{a.s.} 0$. Thus, we rewrite the left hand side of the claim as

$$\begin{aligned}\sum_{i=1}^n v_i \psi_E(\lambda_i) &= \sum_{i=1}^n v_i \psi_H(\lambda_i) \frac{\psi_E(\lambda_i)}{\psi_H(\lambda_i)} = \sum_{i=1}^n v_i (\lambda_i^2/8)(1+R_i) \\ &= \sum_{i=1}^n v_i \cdot \frac{2\log(2/\alpha)}{8\widehat{n}\sigma_{i-1}^2} \cdot (1+R_i) \\ &= \sum_{i=1}^n v_i \cdot \frac{2\log(2/\alpha)}{8n\sigma^2} \cdot (1+R_i') \cdot (1+R_i) \\ &= \sum_{i=1}^n 4(X_i - \widehat{\mu}_{i-1})^2 \cdot \frac{2\log(2/\alpha)}{8n\sigma^2} \cdot (1+R_i+R_i'+R_iR_i').\end{aligned}$$

Defining $R_i'' = R_i + R_i' + R_iR_i'$ for brevity, and noting that $R_i'' \to 0$ almost surely, the above expression becomes

$$\begin{aligned}\sum_{i=1}^n v_i \psi_E(\lambda_i) &= \sum_{i=1}^n (X_i - \widehat{\mu}_{i-1})^2 \cdot \frac{\log(2/\alpha)}{n\sigma^2} \cdot (1+R_i'') \\ &= \frac{\log(2/\alpha)}{\sigma^2} \left[\frac{1}{n} \sum_{i=1}^n (X_i - \widehat{\mu}_{i-1})^2 \cdot (1+R_i'') \right] \\ &= \frac{\log(2/\alpha)}{\sigma^2} \left[\underbrace{\frac{1}{n} \sum_{i=1}^n (X_i - \widehat{\mu}_{i-1})^2}_{\xrightarrow{a.s.} \sigma^2 \text{ by Lemma 6}} + \underbrace{\frac{1}{n} \sum_{i=1}^n (X_i - \widehat{\mu}_{i-1})^2 R_i''}_{\xrightarrow{a.s.} 0 \text{ by Lemma 5}} \right] \xrightarrow{a.s.} \log(2/\alpha),\end{aligned}$$

which completes the proof of Lemma 7. □

Now, consider the denominator in (66).

LEMMA 8. *Continuing with the same notation,*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \lambda_i \xrightarrow{a.s.} \sqrt{\frac{2\log(2/\alpha)}{\sigma^2}}.$$

PROOF. Let R'_i be as in (68). Then,

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\lambda_i &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sqrt{\frac{2\log(2/\alpha)}{n\widehat{\sigma}_{i-1}^2}} \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sqrt{\frac{2\log(2/\alpha)}{n\sigma^2}}\cdot\left(1+R'_i\right) \\ &= \sqrt{\frac{2\log(2/\alpha)}{\sigma^2}}\cdot\underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(1+R'_i\right)}_{\xrightarrow{a.s.}1 \text{ by Lemma 4}} \xrightarrow{a.s.} \sqrt{\frac{2\log(2/\alpha)}{\sigma^2}},\end{aligned}$$

completing the proof of Lemma 8. □

We are now able to combine Lemmas 7 and 8 to prove the main result.

PROPOSITION 9. *Denoting the half-width of* $C_n^{\text{PrPl-EB}(n)}$ *as* W_n*, and assuming the data are drawn iid from a distribution* $Q \in \mathcal{Q}^\mu$ *with variance* σ^2*, we have*

$$\sqrt{n}W_n \;\equiv\; \sqrt{n}\left(\frac{\log(2/\alpha)+\sum_{i=1}^{n}v_i\psi_E(\lambda_i)}{\sum_{i=1}^{n}\lambda_i}\right) \xrightarrow{a.s.} \sigma\sqrt{2\log(2/\alpha)}. \tag{69}$$

Thus, the width is asymptotically proportional to the standard deviation.

PROOF. By direct rearrangement of the left hand side, we see that

$$\begin{aligned}\sqrt{n}\left(\frac{\log(2/\alpha)+\sum_{i=1}^{n}v_i\psi_E(\lambda_i)}{\sum_{i=1}^{n}\lambda_i}\right) &= \frac{\log(2/\alpha)+\sum_{i=1}^{n}v_i\psi_E(\lambda_i)}{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\lambda_i} \\ &\xrightarrow{a.s.} \frac{\log(2/\alpha)+\log(2/\alpha)}{\sigma^{-1}\sqrt{2\log(2/\alpha)}} = \sigma\sqrt{2\log(2/\alpha)},\end{aligned}$$

which completes the proof of Proposition 9. □

E.4. aGRAPA sublevel sets need not be intervals: a worst-case example

In the proof of Theorem 3, we demonstrated that the hedged capital process with predictable plug-in bets yielded convex confidence sets, making their construction more practical. However, this proof was made simple by taking advantage of the fact that the sequences before truncation $(\dot{\lambda}_t^+)_{t=1}^\infty$ and $(\dot{\lambda}_t^-)_{t=1}^\infty$ did not depend on $m \in [0,1]$. This raises the natural question, of whether there are betting-based confidence sets which are nonconvex when these sequences depend on m. Here, we provide a (somewhat pathological) example of the aGRAPA process with nonconvex sublevel sets.

Consider the aGRAPA bets,

$$\lambda_t^{\text{aGRAPA}} := \frac{\widehat{\mu}_{t-1}-m}{\widehat{\sigma}_{t-1}^2+(\widehat{\mu}_{t-1}-m)^2} \text{ where } \widehat{\mu}_t := \frac{1/2+\sum_{i=1}^{t}X_i}{t+1},\ \widehat{\sigma}_t^2 := \frac{1/20+\sum_{i=1}^{t}(X_i-\widehat{\mu}_i)^2}{t+1}. \tag{70}$$

Furthermore, suppose that the observed variables are $X_1 = X_2 = 0$. Then it can be verified that

$$\begin{aligned}\mathcal{K}_2^{\mathrm{aGRAPA}}(m) &= \left(1 + \lambda_1^{\mathrm{aGRAPA}}(X_1 - m)\right)\left(1 + \lambda_2^{\mathrm{aGRAPA}}(X_2 - m)\right) \\ &= \left(1 + \frac{1/2 - m}{1/20 + (1/2 - m)^2}(-m)\right)\left(1 + \frac{1/4 - m}{0.05625 + (1/4 - m)^2}(-m)\right),\end{aligned}$$

which does not yield convex sublevel sets. For example, $\mathcal{K}_2^{\mathrm{aGRAPA}}(0.08) < 0.85$ and $\mathcal{K}_2^{\mathrm{aGRAPA}}(0.4) < 0.85$ but $\mathcal{K}_2^{\mathrm{aGRAPA}}(0.03) > 0.85$. In particular, the sublevel set,

$$\left\{m \in [0,1] : \mathcal{K}_2^{\mathrm{aGRAPA}}(m) < 0.85\right\}$$

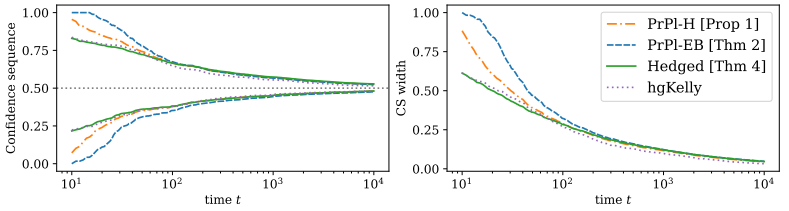
is not convex. In our experience, however, situations like the above do not arise frequently. In fact, we needed to actively search for these examples and use a rather small "prior" variance of $1/20$ which we would not use in practice. Furthermore, the sublevel set given above is at the 0.85 level while confidence sets are compared against $1/\alpha$ which is always larger than 1 and typically larger than 10. We believe that it may be possible to restrict $(\lambda_t^{\mathrm{aGRAPA}})_{t=1}^{\infty}$ and/or the confidence level, $\alpha \in (0,1)$ in some way so that the resulting confidence sets are convex. One reason to suspect that this may be possible is because of the intimate relationship between $\lambda_t^{\mathrm{aGRAPA}}$, $\lambda_t^{\mathrm{GRAPA}}$, and the optimal hindsight bets, λ^{HS}. Specifically, we show in Section E.6 that the optimal hindsight capital $\mathcal{K}_t^{\mathrm{HS}}$ is exactly the empirical likelihood ratio (Owen, 2001) which is known to generate convex confidence sets for the mean (Hall and La Scala, 1990). We leave this question as a direction for future work.

E.5. Betting confidence sequences for non-iid data

The CSs presented in this paper are valid under the assumption that each observation is bounded in $[0,1]$ with conditional mean μ. That is, we require that $X_1, X_2, \ldots$ are $[0,1]$-valued with $\mathbb{E}(X_t \mid \mathcal{F}_{t-1}) = \mu$ for each t, which includes familiar regimes such as independent and identically-distributed (iid) data from some common distribution P with mean μ. Despite the generality of our results, we made matters simpler by focusing the simulations in Section C on the iid setting. For the sake of completeness, we present a simulation to examine the behavior of our CSs in the presence of some non-iid data.

In this setup, we draw the first several hundred or thousand observations independently from a Beta(10, 10) — a distribution whose mean is $1/2$ but whose variance is small (≈ 0.012) — while the remaining observations are independently drawn from a Bernoulli($1/2$) whose mean is also $1/2$ but with a maximal variance of $1/4$. We chose to start the data off with low-variance observations in an attempt to "trick" our betting strategies into adapting to the wrong variance. Empirically, we find that the hedged capital (Theorem 3) and ConBo (Corollary 1) CSs start off strong, adapting to the small variance of a Beta(10, 10). After several Bernoulli($1/2$) observations, the CSs remain tight, but seem to shrink less rapidly. Nevertheless, we find that the hedged capital and ConBo CSs greatly outperform the Hoeffding (Proposition 1) and empirical Bernstein (Theorem 2) predictable plug-in CSs (see Figure 20). Regardless of empirical performance, all methods considered produce *valid* CSs for μ.

250 observations from Beta(10, 10), followed by all Bernoulli(1/2)



2500 observations from Beta(10, 10), followed by all Bernoulli(1/2)

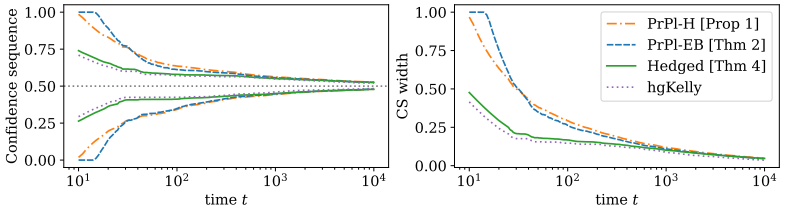


Figure 20. CSs for the true mean $\mu = 1/2$ for non-iid data. In top pair of plots, the first 250 observations were independently drawn from a Beta(10, 10) while the subsequent observations are drawn from a Bernoulli(1/2). The bottom pair of plots is similar, but with 2500 initial draws from a Beta(10, 10) instead of 250. In both cases, the betting-based CSs (Hedged and ConBo) tend to outperform those based on supermartingales.

E.6. Owen's empirical likelihood ratio and Mykland's dual likelihood ratio

Let $x_1, \ldots, x_t \in [0,1]$ and recall the optimal hindsight capital process $\mathcal{K}_t^{\mathrm{HS}}(m)$,

$$\mathcal{K}_t^{\mathrm{HS}}(m) := \prod_{i=1}^{t}(1 + \lambda^{\mathrm{HS}}(x_i - m)) \quad \text{where } \lambda^{\mathrm{HS}} \text{ solves} \quad \sum_{i=1}^{t} \frac{x_i - m}{1 + \lambda^{\mathrm{HS}}(x_i - m)} = 0.$$

Now, let $\mathcal{Q}^m \equiv \mathcal{Q}^m(x_1^t)$ be the collection of discrete probability measures with support $\{x_1, \ldots, x_t\}$ and mean m. Let $\mathcal{Q} \equiv \mathcal{Q}(x_1^t) := \bigcup_{m \in [0,1]} \mathcal{Q}^m$ and define the empirical likelihood ratio (Owen, 2001),

$$\mathrm{EL}_t(m) := \frac{\sup_{Q \in \mathcal{Q}} \prod_{i=1}^{t} Q(x_i)}{\sup_{Q \in \mathcal{Q}^m} \prod_{i=1}^{t} Q(x_i)}.$$

Owen (2001) showed that the numerator equals $(1/t)^t$ and the denominator equals

$$\prod_{i=1}^{t}(1 + \lambda^{\mathrm{EL}}(x_i - m))^{-1} \quad \text{where } \lambda^{\mathrm{EL}} \text{ solves} \quad \sum_{i=1}^{t} \frac{x_i - m}{1 + \lambda^{\mathrm{EL}}(x_i - m)} = 0.$$

Notice that the above product is exactly the reciprocal of $\mathcal{K}_t^{\mathrm{HS}}$ and that $\lambda^{\mathrm{EL}} = \lambda^{\mathrm{HS}}$. Therefore for each $m \in [0,1]$,

$$\mathrm{EL}_t(m) = (1/t)^t \mathcal{K}_t^{\mathrm{HS}}(m).$$

Furthermore, given the connection between the empirical and dual likelihood ratios for independent data (Mykland, 1995), the hindsight capital process is also proportional to the dual likelihood ratio in this case.

F. An extended history of betting and its applications

(This is an expanded version of Section 6 and Figure 9.)

The use of betting-related ideas in probability, statistics, optimization, finance and machine learning has evolved in many different parallel threads, emanating from different influential early works and thus having different roots and evolutions. Since these threads have had little interaction for many decades now, we consider it worthwhile to mention them in some detail. Two notes of caution:

- We anticipate missing some authors and works in our broad strokes below, but a thorough coverage would be better suited to a longer survey paper on the topic. For example, we entirely skip the field of mathematical finance, since betting is literally a foundation of the entire field (and theoretical and applied progress on martingales, betting strategies, and related topics has been phenomenal).
- Many of the authors listed below have used the language of betting in their works explicitly, but others have not — and may even prefer (or have preferred) *not* to do so. Thus, our references should be treated with a pinch of salt, as some connections that we draw to betting may be more apparent in hindsight (to us) than foresight (to the authors).

If we had to pick the most critical early authors without whom our work would have been impossible, it would be Ville, Wald, Kelly and Robbins; later influences on us have been via Lai, Cover, Shafer, Vovk, Grunwald and the second author's own earlier works (Howard et al., 2020, 2021). These authors stand out below.

Probability. Ville's 1939 PhD thesis (Ville, 1939) contained an important and rather remarkable result of its time that connected measure-theoretic probability with betting, and indeed brought the very notion of a martingale into probability theory. In brief, Ville proved that for every event of measure zero, there exists a betting strategy for which a gambler's wealth process (a nonnegative martingale) grows to infinity if that event occurs. For example, the strong law of large numbers (SLLN) and the law of the iterated logarithm (LIL) are two classic measure-theoretic statements that occur on all sequences of observations, except for a null set according to some underlying probability measure (where the two null sets for the two laws are different). Ville proved that it is possible to bet on the next outcome such that if the LIL were false for that particular sequence of observations, then the gambler's wealth would grow in an unbounded fashion.

Doob's monumental papers and book Doob (1953) in the following decades stripped martingales of their betting roots and presented them as some of the most powerful tools of measure-theoretic probability theory, with applications to many other branches of mathematics. (However, betting could be viewed as instances of "Doob's martingale transform".) These betting roots were revived in the 1960s with the renewed interest in algorithmic definitions of randomness, due to Kolmogorov, Martin-Löf (1966) and many others.

More recently, Shafer and Vovk (2001, 2019) have produced two seminal books that aim bring betting and martingales to the front and center of probability and finance,

aiming to derive much (if not all) of probability theory from purely game-theoretic principles based on betting strategies. The product martingale wealth process that appears in our work also appears in theirs (indeed, it is a fundamental process), but Shafer and Vovk did not explore the topics in our paper (confidence sequences, explicit computationally efficient betting strategies, sampling without replacement, thorough numerical simulations, and so on). Indeed, their book has a thorough treatment of probability and finance, but with respect to statistical inference, there is little explicit methodology for practice. Perhaps they were aware of such a statistical utility, but they did not explicitly recognize or demonstrate the excellent power of betting in practice (when properly developed) for problems such as ours.

Statistical inference. Using the power of hindsight, we now know that Wald's influential work on the sequential probability ratio test was implicitly based on martingale techniques Wald (1945). Wald derived many fundamental results that he required from scratch without having the general language that was being set up by Doob in parallel to his work. In the case of testing a simple null $H_0 : \theta = \theta^*$ against a composite alternative $H_0 : \theta \neq \theta^*$, Wald (1945, Eq (10:10)) suggests forming the likelihood ratio process $\prod_{i=1}^n f_{\theta_{i-1}}(X_i) / \prod_{i=1}^n f_{\theta^*}(X_i)$, where θ_{i-1} is a mapping from $X_1, \ldots, X_{i-1}$ to Θ; in other words, θ_{i-1} is predictable. In the language of our paper, this is a predictable plug-in, and the first appearance of betting-like ideas in the statistical literature. However, beyond this passing equation in a parametric setup, the idea appears to have lain dormant.

Robbins (along with students and colleagues Siegmund, Darling, and Lai) quickly realized the power of Wald's and Ville's ideas as well as martingales more generally, and pursued a rather broad agenda around sequential testing and estimation, including the introduction and extensive study of confidence sequences and the method of mixtures (Darling and Robbins, 1967c,a,b; Robbins and Siegmund, 1968, 1969, 1970, 1972, 1974; Lai, 1976). Robbins and Siegmund also analyzed Wald's "betting" test, and proved in some generality that its behavior is similar to a mixture likelihood ratio test (Robbins and Siegmund, 1974, Section 6). Most of Wald's and Robbins' work was parametric, but Robbins did explicitly study the sub-Gaussian setting in some detail (Robbins, 1970). Building on a vast literature of Chernoff-style concentration inequalities that exploded after Robbins' time, Howard et al. (2020, 2021) recently extended mixture methods of Robbins to derive confidence sequences under a large class of nonparametric settings using exponential supermartingales. Howard et al. (2020, 2021) recognized Wald's betting idea, but did not develop it nonparametrically beyond a brief mention in the paper as a direction for future work. The current work takes this natural next step in some thorough detail.

Information and coding theory. Soon after the seminal work of Shannon (1948), another researcher at AT&T Bell Labs, John Larry Kelly Jr. wrote a paper titled "A New Interpretation of Information Rate" which explicitly connected betting with the new field of information theory, complementing the work of Shannon (Kelly Jr, 1956). In short, he proved that it is possible to bet on the symbols in a communication channel at odds consistent with their probabilities in order to have a gambler's

wealth grow exponentially, with the exponent equaling the rate of transmission over the channel. More explicitly, given a sequence of Bernoulli random variables with probability $p > 1/2$, Kelly proved that betting a $(2p - 1)$ fraction of your current wealth on the next outcome being 1 is the unique strategy that maximizes the expected log wealth of the gambler.

When the probability p changes at each step in an unknown manner, the "universal coding" work of Krichevsky and Trofimov (1981) showed that a mixture method involving the Jeffreys prior and maximum likelihood can achieve nearly the optimal wealth in hindsight, with the expected log wealth of their strategy only being worse than the optimal oracle log-wealth by a factor that is logarithmic in the number of rounds; these observations work for any discrete alphabet, not just a binary. 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), that today forms a standard textbook topic in information theory.

There are other parts of information/coding theory that could be seen as related in some ways to betting through the use of (what are now called) e-variables: these include the topics of prequential model selection and minimum description length; see works by Rissanen (1984, 1998), Dawid (1984, 1997), Grünwald (2007); Grünwald et al. (2019), Li (1999) and references therein.

Online learning and sequential prediction under log loss. In the 1990s, the problems studied by Krichevsky, Trofimov, and Cover continued to be extended — often dropping the information theoretic context — under the title of sequential prediction under the logarithmic loss. In the active subfield of online learning, the previous results were effectively "regret bounds" against potentially adversarial sequences of observations, with a chapter devoted to the problem in the book on prediction, learning and games by Cesa-Bianchi and Lugosi (2006). More recently, Orabona and colleagues such as Pal and Jun have found powerful implications of these ideas in deriving parameter-free algorithms for online convex optimization (Orabona and Pal, 2016; Orabona and Tommasi, 2017; Jun et al., 2017; Jun and Orabona, 2019).

Rakhlin and Sridharan (2017) found that deterministic regret inequalities can be used to derive concentration inequalities for martingales, connecting the two rich fields. Later, Jun and Orabona (2019) also derive concentration inequalities using their betting-based regret bounds, with explicit bounds derived in the sub-Gaussian and bounded settings. However, because regret bounds could be tight in rate but are typically loose in constants, the resulting concentration inequalities are not tight in practice. Thus, we view this line of work as important and complementary to our explorations, which are different in their motivation, derivation and practicality.

Typically, none of these lines of literature have cited the others. For example, the important paper of Rakhlin and Sridharan (2017) does not mention the work of Ville, Wald or Robbins, or even of Vovk and Shafer. Similarly, despite the books of Shafer and Vovk having a wonderful coverage of the history of probability and martingales stemming back hundreds of years, even their recent 2019 book (Shafer and Vovk, 2019) does not cite the coding theory and online learning literature very much,

including the works of Orabona and coauthors (Orabona and Pal, 2016; Orabona and Tommasi, 2017; Jun et al., 2017; Cutkosky and Orabona, 2018; Jun and Orabona, 2019), Krichevsky and Trofimov (1981), or Rakhlin and Sridharan (2017). Recent work of Orabona and colleagues also in turn has no mention of the books of Shafer and Vovk (2001, 2019), or works of Ville, Wald, Robbins, Howard, their coauthors and other recent authors. The work of Howard et al. (2020, 2021) does cite the Wald and Robbins literatures, as well as the books of Shafer and Vovk and pioneering work of Ville, but does not form connections to information/coding theory nor to online learning. The excellent book of Cesa-Bianchi and Lugosi (2006) does not cite Ville, the seminal martingale works of Robbins, or the 2001 book by Shafer and Vovk. ‖

The reason for the lack of intersection of these parallel threads is likely manifold, and definitely far from malicious: (a) these works were and continue to be published in different literatures, (b) these works had different goals in mind, meaning that they were addressing different problems and often using different techniques, (c) our understanding of these literatures and their relationships is constantly evolving and far from complete; it is likely that no author has a command over all these parallel literatures, and indeed this should not be expected.

In the preface of their 2006 book, Cesa-Bianchi and Lugosi write

> Prediction of individual sequences, the main theme of this book, has been studied in various fields, such as statistical decision theory, information theory, game theory, machine learning, and mathematical finance. Early appearances of the problem go back as far as the 1950s, with the pioneering work of Blackwell, Hannan, and others. Even though the focus of investigation varied across these fields, some of the main principles have been discovered independently. Evolution of ideas remained parallel for quite some time. As each community developed its own vocabulary, communication became difficult. By the mid-1990s, however, it became clear that researchers of the different fields had a lot to teach each other. When we decided to write this book, in 2001, one of our main purposes was to investigate these connections and help ideas circulate more fluently. In retrospect, we now realize that the interplay among these many fields is far richer than we suspected. ... Today, several hundreds of pages later, we still feel there remains a lot to discover. This book just shows the first steps of some largely unexplored paths. We invite the reader to join us in finding out where these paths lead and where they connect.

Thus it is clear that Cesa-Bianchi and Lugosi already foresaw that there were many connections between the fields that have been unstated, underappreciated, undiscovered and underutilized. The connections we briefly point out above between these literatures, both historical and modern, are themselves new in their own right (not existing in any of the aforementioned books or papers) and may be considered a small contribution of this paper. A more thorough investigation of these connections may be the topic of a future survey paper, or indeed, a book on these topics.

‖Authors like like Rissanen (1984, 1998) and Dawid (1984, 1997) are not cited in most of these works, perhaps because the connections of their works to betting are indirect.

Supplemental References

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.

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.

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.

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.

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, 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.

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.

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.

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.

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.

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, 1974.

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.

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

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

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

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.