The Annals of Statistics
2021, Vol. 49, No. 2, 1055–1080
https://doi.org/10.1214/20-AOS1991

TIME-UNIFORM, NONPARAMETRIC, NONASYMPTOTIC CONFIDENCE SEQUENCES

BY STEVEN R. HOWARD[1,*], AADITYA RAMDAS[2], JON MCAULIFFE[1,†] AND JASJEET SEKHON[3]

[1]*Department of Statistics, University of California, Berkeley,* [*]*stevehoward@berkeley.edu;* [†]*jonmcauliffe@berkeley.edu*
[2]*Departments of Statistics and Machine Learning, Carnegie Mellon University, aramdas@stat.cmu.edu*
[3]*Department of Statistics and Data Science, Yale University, sekhon@berkeley.edu*

A confidence sequence is a sequence of confidence intervals that is uniformly valid over an unbounded time horizon. Our work develops confidence sequences whose widths go to zero, with nonasymptotic coverage guarantees under nonparametric conditions. We draw connections between the Cramér–Chernoff method for exponential concentration, the law of the iterated logarithm (LIL) and the sequential probability ratio test—our confidence sequences are time-uniform extensions of the first; provide tight, nonasymptotic characterizations of the second; and generalize the third to nonparametric settings, including sub-Gaussian and Bernstein conditions, self-normalized processes and matrix martingales. We illustrate the generality of our proof techniques by deriving an empirical-Bernstein bound growing at a LIL rate, as well as a novel upper LIL for the maximum eigenvalue of a sum of random matrices. Finally, we apply our methods to covariance matrix estimation and to estimation of sample average treatment effect under the Neyman–Rubin potential outcomes model.

1. Introduction.

It has become standard practice for organizations with online presence to run large-scale randomized experiments, or “A/B tests,” to improve product performance and user experience. Such experiments are inherently sequential: visitors arrive in a stream and outcomes are typically observed quickly relative to the duration of the test. Results are often monitored continuously using inferential methods that assume a fixed sample, despite the known problem that such monitoring inflates Type I error substantially [1, 8]. Furthermore, most A/B tests are run with little formal planning and fluid decision-making, compared to clinical trials or industrial quality control, the traditional applications of sequential analysis.

This paper presents methods for deriving *confidence sequences* as a flexible tool for inference in sequential experiments [12, 32, 43]. For $\alpha \in (0, 1)$, a $(1-\alpha)$-confidence sequence is a sequence of confidence sets $(\mathrm{CI}_t)_{t=1}^{\infty}$, typically intervals $\mathrm{CI}_t = (L_t, U_t) \subseteq \mathbb{R}$, satisfying a uniform coverage guarantee: after observing the tth unit, we calculate an updated confidence set CI_t for the unknown quantity of interest θ_t, with the uniform coverage property

$$\mathbb{P}(\forall t \geq 1 : \theta_t \in \mathrm{CI}_t) \geq 1 - \alpha. \tag{1.1}$$

With only a uniform lower bound (L_t), that is, if $U_t \equiv \infty$, we have a *lower confidence sequence*. Likewise, if $L_t \equiv -\infty$ we have an *upper confidence sequence* given by (U_t). Theorems 1 to 3 and Lemma 2 are our key tools for constructing confidence sequences. All build upon the general framework for uniform exponential concentration introduced in Howard et al. [25], which means our techniques apply in diverse settings: scalar, matrix and Banach-space-valued observations, with possibly unbounded support; self-normalized bounds applicable to observations satisfying weak moment or symmetry conditions; and continuous-time

Received February 2019; revised May 2020.
MSC2020 subject classifications. Primary 62L12, 62G05; secondary 60G42, 60B20.
Key words and phrases. Confidence sequence, finite LIL bound, empirical-Bernstein bound, sequential probability ratio test, matrix concentration, potential outcomes.

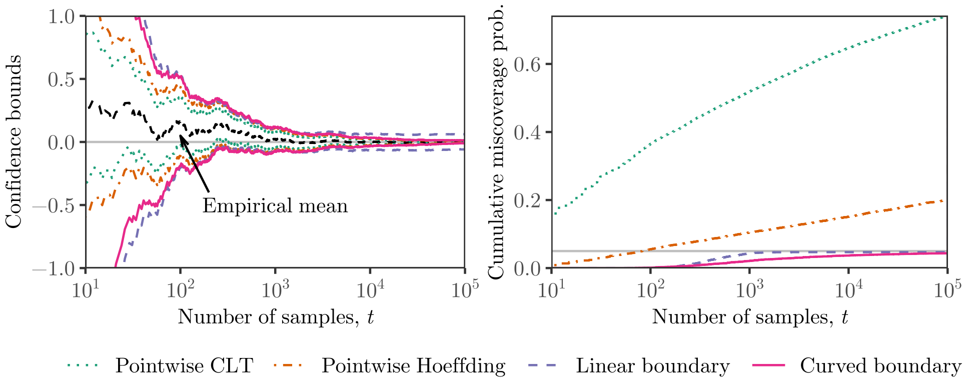

FIG. 1. *Left panel shows* 95% *pointwise confidence intervals and uniform confidence sequences for the mean of a Rademacher random variable*, *using one simulation of* 100,000 *i.i.d. draws. Right panel shows cumulative chance of miscoverage based on* 10,000 *replications*; *flat grey line shows the nominal target level* 0.05. *The CLT intervals are asymptotically pointwise valid* (*these are similar to the exact binomial confidence intervals*, *which are nonasymptotically pointwise valid*). *The pointwise Hoeffding intervals are nonasymptotically pointwise valid. The confidence sequence based on a linear boundary*, *as in Lemma* 1, *is valid uniformly over time and nonasymptotically*, *but does not shrink to zero width. Finally*, *the confidence sequence based on a curved boundary is valid uniformly and nonasymptotically*, *while also shrinking towards zero width*; *here we use the two-sided normal mixture boundary*, (3.7), *qualitatively similar to the stitched bound* (1.2).

scalar martingales. Our methods allow for flexible control of the "shape" of the confidence sequence, that is, how the sequence of intervals shrinks in width over time. As a simple example, given a sequence of i.i.d. observations $(X_t)_{t=1}^{\infty}$ from a 1-sub-Gaussian distribution whose mean μ we would like to estimate, Theorem 1 yields the following $(1-\alpha)$-confidence sequence for μ, a special case of the more general bound (3.3):

$$\frac{\sum_{i=1}^{t} X_i}{t} \pm 1.7\sqrt{\frac{\log\log(2t) + 0.72\log(5.2/\alpha)}{t}}. \tag{1.2}$$

The $\mathcal{O}(\sqrt{t^{-1}\log\log t})$ asymptotic rate of this bound matches the lower bound implied by the law of the iterated logarithm (LIL), and nonasymptotic bounds of this form are called *finite LIL bounds* [29]. We develop confidence sequences that possess the following properties:

(P1) *Nonasymptotic and nonparametric*: our confidence sequences offer coverage at all sample sizes without exact distributional assumptions or asymptotic approximations.

(P2) *Unbounded sample size*: we do not require a final sample size to be chosen ahead of time. They may be tuned for a planned sample size but always permit additional sampling.

(P3) *Arbitrary stopping rules*: we make no assumptions on the stopping rule used by an experimenter to decide when to end the experiment, or when to act on certain inferences.

(P4) *Asymptotically zero width*: the interval widths of our confidence sequences shrink toward zero at a $1/\sqrt{t}$ rate, ignoring log factors, just as with pointwise confidence intervals.

These properties give us strong guarantees and broad applicability. An experimenter may always choose to gather more samples, and may stop at any time according to any rule—the resulting inferential guarantees hold under the stated assumptions without any approximations. Of course, this flexibility comes with a cost: our intervals are wider than those that rely on asymptotics or make stronger assumptions, for example, a known stopping rule. Typical, fixed-sample confidence intervals derived from the central limit theorem do not satisfy any of (P1)–(P3), and accommodating any one property necessitates wider intervals; we illustrate this in Figure 1. It is perhaps surprising that these four properties come at a numerical cost of less than doubling the fixed-sample, asymptotic interval width—the discrete mixture bound illustrated in Figure S2 in the Supplementary Material [26] stays within a factor of two of the fixed-sample CLT bounds over five orders of magnitude in time.

1.1. *Related work.* The idea of a confidence sequence goes back at least to Darling and Robbins [12]. They are called *repeated confidence intervals* by Jennison and Turnbull [31, 32] (with a focus on finite time horizons) and *always-valid confidence intervals* by Johari, Pekelis and Walsh [35]. They are sometimes labeled *anytime confidence intervals* in the machine learning literature [28].

Prior work on sequential inference is often phrased in terms of a sequential hypothesis test, defined as a stopping rule and an accept/reject decision variable, or in terms of an always-valid p-value [35]. In Section 6, we discuss the duality between confidence sequences, sequential hypothesis tests, and always-valid p-values. We show in Lemma 3 that definition (1.1) is equivalent to requiring $\mathbb{P}(\theta_\tau \in \mathrm{CI}_\tau) \geq 1 - \alpha$ for all stopping times τ, or even for all random times τ, not necessarily stopping times. Hence the choice of definition (1.1) over related definitions in the literature is one of convenience.

Recent interest in confidence sequences has come from the literature on best-arm identification with fixed confidence for multi-armed bandit problems. Garivier [20], Jamieson et al. [29], Kaufmann, Cappé and Garivier [37] and Zhao et al. [77] present methods satisfying properties (P1)–(P4) for independent, sub-Gaussian observations. Our results are sharper and more general, and our Bernstein confidence sequence scales with the true variance in nonparametric settings. Confidence sequences are a key ingredient in best-arm selection algorithms [30] and related methods for sequential testing with multiple comparisons [28, 49, 76]. Our results improve and generalize such methods.

Maurer and Pontil [50] and Audibert, Munos and Szepesvári [3] prove empirical-Bernstein bounds for fixed times or finite time horizons. Our empirical-Bernstein bound holds uniformly over infinite time. Balsubramani [5] takes a different approach to deriving confidence sequences satisfying properties (P1)–(P4) by lower bounding a mixture martingale. This work was extended in Balsubramani and Ramdas [6] to an empirical-Bernstein bound, the only infinite-horizon, empirical-Bernstein confidence sequence we are aware of in prior work. Our result removes a multiplicative prefactor and yields sharper bounds. We emphasize that our proof technique is quite different from all three of these existing empirical Bernstein bounds; see Appendix A.8.

The simplest confidence sequence satisfying properties (P1)–(P3) follows by inverting a suitably formulated sequential probability ratio test (SPRT, [73]), such as in Section 3.6 of Howard et al. [25]. Wald worked in a parametric setting, though it is known that the normal SPRT depends only on sub-Gaussianity (e.g., Robbins [55]). The resulting confidence sequence does not shrink toward zero width as $t \to \infty$ (property P4), a problem which stems from the choice of a single point alternative λ. Numerous extensions have been developed to remedy this defect, and our work is most closely tied to two approaches. First, in the method of mixtures, one replaces the likelihood ratio with a mixture $\int \prod_i [f_\lambda(X_i)/f_0(X_i)] \, \mathrm{d}F(\lambda)$, which is still a martingale [5, 7, 14, 16, 38, 41, 55, 57, 58, 72, 73]. Second, epoch-based analyses choose a sequence of point alternatives $\lambda_1, \lambda_2, \ldots$ approaching the null value, with corresponding error probabilities $\alpha_1, \alpha_2, \ldots$ approaching zero so that a union bound yields the desired error control [13, 37, 56].

The literature on self-normalized bounds makes extensive use of the method of mixtures, sometimes called pseudo-maximization [15–18, 20]; these works introduced the idea of using a mixture to bound a quantity with a random intrinsic time V_t. These results are mostly given for fixed samples or finite time horizon, though de la Peña, Klass and Lai [15], equation (4.20), includes an infinite-horizon curve-crossing bound. Lai [41] treats confidence sequences for the parameter of an exponential family using mixture techniques similar to those of Section 3.2. Like most work on the method of mixtures, Lai's work focused on the parametric setting (which we discuss in Section 4.4), while we focus on the application of mixture bounds to nonparametric settings.

Johari et al. [34] adopt the mixture approach for a commercial A/B testing platform, where properties (P2) and (P3) are critical to provide an "off-the-shelf" solution for a variety of clients. Their application relies on asymptotics which lack rigorous justification. In Section 4.2, we give nonasymptotic justification for a similar confidence sequence under a finite-sample randomization inference model, and in Section 5 we demonstrate how our methods control Type I error in situations where asymptotics fail.

1.2. *Outline.* We organize our results using the sub-Gaussian, sub-gamma, sub-Bernoulli, sub-Poisson and subexponential settings defined in Section 2.

1. The *stitching* method gives new closed-form sub-Gaussian or sub-gamma boundaries (Theorem 1). Our sub-gamma treatment extends prior sub-Gaussian work to cover any martingale whose increments have finite moment-generating function in a neighborhood of zero; see Proposition 1. Our proof is transparent and flexible, accommodating a variety of boundary shapes, including those growing at the rate $\mathcal{O}(\sqrt{t \log\log t})$ with a focus on tight constants, though we do not recommend this bound in practice unless closed-form simplicity is vital.
2. *Conjugate mixtures* give one- and two-sided boundaries for the sub-Bernoulli, sub-Gaussian, sub-Poisson and subexponential cases (Section 3.2) which avoid approximations made for analytical convenience. The sub-Gaussian boundaries are unimprovable without further assumptions (Section 3.6). These boundaries include a common tuning parameter which is critical in practice and we discuss why their $\mathcal{O}(\sqrt{t \log t})$ growth rate may be preferable to the slower $\mathcal{O}(\sqrt{t \log\log t})$ rate (Section 3.5).
3. *Discrete mixtures* facilitate numerical computation of boundaries with a great deal of flexibility, at the cost of slightly more involved computations (Theorem 2). Like conjugate mixture boundaries, these boundaries avoid unnecessary approximations and are unimprovable in the sub-Gaussian case.
4. Finally, for sub-Gaussian processes, the *inverted stitching* method (Theorem 3) gives numerical upper bounds on the crossing probability of *any* increasing, strictly concave boundary over a limited time range. We show that any such boundary yields a uniform upper tail inequality over a finite horizon, and compute its crossing probability.

Building on this foundation, we present a a state-of-the-art empirical-Bernstein bound (Theorem 4) for any sequence of bounded observations using a new self-normalization proof technique. We illustrate our methods with two novel applications: the nonasymptotic, sequential estimation of average treatment effect in the Neyman–Rubin potential outcomes model (Section 4.2), and the derivation of uniform matrix bounds and covariance matrix confidence sequences (Corollary 3 and Section 4.3). We give simulation results in Section 5. Section 6 discusses the relationship of our work to existing concepts of sequential testing. Proofs of main results are in Appendix A, with others deferred to Appendix C.

2. Preliminaries: Linear boundaries. Given a sequence of real-valued observations $(X_t)_{t=1}^{\infty}$, suppose we wish to estimate the average conditional expectation $\mu_t := t^{-1} \times \sum_{i=1}^{t} \mathbb{E}_{i-1} X_i$ at each time t using the sample mean $\bar{X}_t := t^{-1} \sum_{i=1}^{t} X_i$; here we assume an underlying filtration $(\mathcal{F}_t)_{t=1}^{\infty}$ to which (X_t) is adapted, and $\mathbb{E}_t$ denotes expectation conditional on $\mathcal{F}_t$. Let $S_t := \sum_{i=1}^{t}(X_i - \mathbb{E}_{i-1} X_i)$, the zero-mean deviation of our sample sum from its estimand at time t. Given $\alpha \in (0, 1)$, suppose we can construct a uniform upper tail bound $u_\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ satisfying

$$\mathbb{P}(\exists t \geq 1 : S_t \geq u_\alpha(V_t)) \leq \alpha \tag{2.1}$$

for some adapted, real-valued *intrinsic time* process $(V_t)_{t=1}^{\infty}$, an appropriate time scale to measure the (squared) deviations of (S_t). This uniform upper bound on the centered sum (S_t) yields a lower confidence sequence for (μ_t) with radius $t^{-1}u_\alpha(V_t)$:

$$\mathbb{P}(\forall t \geq 1 : \bar{X}_t - t^{-1}u_\alpha(V_t) \leq \mu_t) \geq 1 - \alpha.$$

Note that an assumption on the upper tail of (S_t) yields a lower confidence sequence for (μ_t); a corresponding assumption on the lower tail of (S_t) yields an upper confidence sequence for (μ_t). In this paper, we formally focus on upper tail bounds, from which lower tail bounds can be derived by examining $(-S_t)$ in place of (S_t). In general, the left and right tails of (S_t) may behave differently and require different sets of assumptions, so that our upper and lower confidence sequences may have different forms. Regardless, we can always combine upper and lower confidence sequences using a union bound to obtain a two-sided confidence sequence (1.1).

When the (X_t) are independent with common mean μ, the resulting confidence sequence estimates μ, but the setup requires neither independence nor a common mean. In general, the estimand μ_t may be changing at each time t; Section 4.2 gives an application to causal inference in which this changing estimand is useful. In principle, μ_t may also be random, although none of our applications involve random μ_t.

To construct uniform boundaries u_α satisfying inequality (2.1), we build upon the following general condition [25], Definition 1.

DEFINITION 1 (Sub-ψ condition). Let $(S_t)_{t=0}^{\infty}$, $(V_t)_{t=0}^{\infty}$ be real-valued processes adapted to an underlying filtration $(\mathcal{F}_t)_{t=0}^{\infty}$ with $S_0 = V_0 = 0$ and $V_t \geq 0$ for all t. For a function $\psi : [0, \lambda_{\max}) \to \mathbb{R}$ and a scalar $l_0 \in [1, \infty)$, we say (S_t) is l_0-*sub*-ψ *with variance process* (V_t) if, for each $\lambda \in [0, \lambda_{\max})$, there exists a supermartingale $(L_t(\lambda))_{t=0}^{\infty}$ w.r.t. $(\mathcal{F}_t)$ such that $\mathbb{E}L_0(\lambda) \leq l_0$ and

$$\exp\{\lambda S_t - \psi(\lambda)V_t\} \leq L_t(\lambda) \quad \text{a.s. for all } t. \tag{2.2}$$

For given ψ and l_0, let $\mathbb{S}_{\psi}^{l_0}$ be the class of pairs of l_0-sub-ψ processes (S_t, V_t):

$$\mathbb{S}_{\psi}^{l_0} := \{(S_t, V_t) : (S_t) \text{ is } l_0\text{-sub-}\psi \text{ with variance process } (V_t)\}. \tag{2.3}$$

When stating that a process is sub-ψ, we typically omit l_0 from our terminology for simplicity. In scalar cases, we always have $l_0 = 1$, while in matrix cases $l_0 = d$, the dimension of the (square) matrices.

Where does Definition 1 come from? The jumping-off point is the martingale method for concentration inequalities ([4, 24, 51]; [54], Section 2.2), itself based on the classical Cramér–Chernoff method ([10, 11]; [9], Section 2.2). The martingale method starts off with an assumption of the form $\mathbb{E}_{t-1}e^{\lambda(X_t - \mathbb{E}_{t-1}X_t)} \leq e^{\psi(\lambda)\sigma_t^2}$ for all $t \geq 1$, $\lambda \in \mathbb{R}$. Then, denoting $S_t := \sum_{i=1}^{t}(X_i - \mathbb{E}_{i-1}X_i)$ and $V_t := \sum_{i=1}^{t}\sigma_i^2$, the process $\exp\{\lambda S_t - \psi(\lambda)V_t\}$ is a supermartingale for each $\lambda \in \mathbb{R}$. Unlike the martingale method assumption, Definition 1 allows the exponential process to be upper bounded by a supermartingale, and it permits (V_t) to be adapted rather than predictable. We also restrict our attention to $\lambda \geq 0$ for one-sided bounds.

Intuitively, the process $\exp\{\lambda S_t - \psi(\lambda)V_t\}$ measures how quickly S_t has grown relative to intrinsic time V_t, and the free parameter λ determines the relative emphasis placed on the tails of the distribution of S_t, that is, on the higher moments. Larger values of λ exaggerate larger movements in S_t, and ψ captures how much we must correspondingly exaggerate V_t. ψ is related to the heavy-tailedness of S_t and the reader may think of it as a cumulant-generating function (CGF, the logarithm of the moment-generating function). For example, suppose (X_t) is a sequence of i.i.d., zero-mean random variables with CGF $\psi(\lambda) := \log \mathbb{E}e^{\lambda X_1}$ which is finite for all $\lambda \in [0, \lambda_{\max})$. Then, setting $V_t := t$, we see that $L_t(\lambda) := \exp\{\lambda S_t - \psi(\lambda)V_t\}$ is itself a martingale, for all $\lambda \in [0, \lambda_{\max})$. Indeed, in all scalar cases, we consider $L_t(\lambda)$ is just equal to $\exp\{\lambda S_t - \psi(\lambda)V_t\}$. See Appendix Tables S3 and S4, drawn from Howard et al. [25], for a catalog of sufficient conditions for a process to be sub-ψ using the five ψ functions defined below. We use many of these conditions in what follows.

We organize our uniform boundaries according to the ψ function used in Definition 1. First recall the Cramér–Chernoff bound: if (X_t) are independent zero-mean with bounded CGF $\log \mathbb{E} e^{\lambda X_t} \le \psi(\lambda)$ for all $t \ge 1$ and $\lambda \in \mathbb{R}$, then writing $S_t = \sum_{i=1}^t X_i$, we have $\mathbb{P}(S_t \ge x) \le e^{-t\psi^\star(x/t)}$ for any $x > 0$, where $\psi^\star$ denotes the Legendre–Fenchel transform of ψ. Equivalently, writing $z_\alpha(t) := t\psi^{\star -1}(t^{-1}\log\alpha^{-1})$, we have $\mathbb{P}(S_t \ge z_\alpha(t)) \le \alpha$ for any fixed t and $\alpha \in (0, 1)$. In other words, the function z_α gives a high-probability upper bound at any fixed time t for *any* sum of independent random variables with CGF bounded by ψ. When we extend this concept to boundaries holding uniformly over time, there is no longer a unique, minimized boundary, and the following definition captures the class of valid boundaries.

DEFINITION 2. Given $\psi : [0, \lambda_{\max}) \to \mathbb{R}$ and $l_0 \ge 1$, a function $u : \mathbb{R} \to \mathbb{R}$ is called an *l_0-sub-ψ uniform boundary* with crossing probability α if

$$\sup_{(S_t, V_t) \in \mathbb{S}_\psi^{l_0}} \mathbb{P}(\exists t \ge 1 : S_t \ge u(V_t)) \le \alpha. \tag{2.4}$$

Although u does depend on the constant l_0 in Definition 1, for simplicity we typically omit this dependence from our notation, writing simply that u is a sub-ψ uniform boundary.

Five particular ψ functions play important roles in our development; below, we take $1/0 = \infty$ in the upper bounds on λ:

- $\psi_{B,g,h}(\lambda) := \frac{1}{gh}\log(\frac{ge^{h\lambda} + he^{-g\lambda}}{g+h})$ on $0 \le \lambda < \infty$, the scaled CGF of a centered random variable (r.v.) supported on two points, $-g$ and h, for some $g, h > 0$, for example, a centered Bernoulli r.v. when $g + h = 1$.
- $\psi_N(\lambda) := \lambda^2/2$ on $0 \le \lambda < \infty$, the CGF of a standard Gaussian r.v.
- $\psi_{P,c}(\lambda) := c^{-2}(e^{c\lambda} - c\lambda - 1)$ on $0 \le \lambda < \infty$ for some scale parameter $c \in \mathbb{R}$, which is the CGF of a centered unit-rate Poisson r.v. when $c = 1$. By taking the limit, we define $\psi_{P,0} = \psi_N$.
- $\psi_{E,c}(\lambda) := c^{-2}(-\log(1 - c\lambda) - c\lambda)$ on $0 \le \lambda < 1/(c \vee 0)$ for some scale $c \in \mathbb{R}$, which is the CGF of a centered unit-rate exponential r.v. when $c = 1$. By taking the limit, we define $\psi_{E,0} = \psi_N$.
- $\psi_{G,c}(\lambda) := \lambda^2/(2(1 - c\lambda))$ on $0 \le \lambda < 1/(c \vee 0)$ (taking $1/0 = \infty$) for some scale parameter $c \in \mathbb{R}$, which we refer to as the sub-gamma case following Boucheron, Lugosi and Massart [9]. This is not the CGF of a gamma r.v. but is a convenient upper bound which also includes the sub-Gaussian case at $c = 0$ and permits analytically tractable results.

One may freely scale ψ by any positive constant and divide V_t by the same constant so that Definition 1 remains satisfied; by convention, we scale all ψ functions above so that $\psi''(0_+) = 1$. When we speak of a *sub-gamma* process (or uniform boundary) with scale parameter c, we mean a sub-$\psi_{G,c}$ process (or uniform boundary), and likewise for other cases. We often write ψ_B, ψ_P, etc., dropping the range and scale parameters from our notation. As we summarize in Figure 2 and detail in Proposition S7, certain general implications hold among sub-ψ boundaries. In particular, any sub-Gaussian boundary can also serve as a sub-Bernoulli boundary; any sub-Poisson boundary serves as a sub-Gaussian or sub-Bernoulli boundary; and, importantly, any sub-gamma or subexponential boundary can serve as a sub-ψ boundary in any of the other four cases. Indeed, a sub-gamma or subexponential boundary applies to many cases of practical interest, as detailed below.

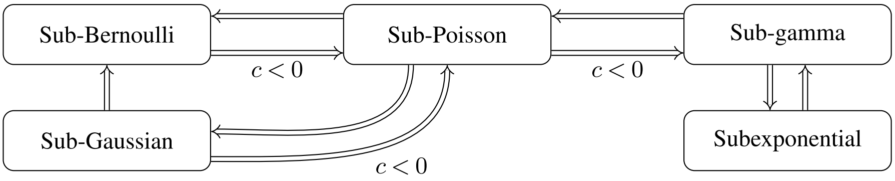

FIG. 2. *Relations among sub-ψ boundaries: each arrow indicates that a sub-ψ boundary at the source node can also serve as a sub-ψ boundary at the destination node, with appropriate modifications to their parameters. Details are in Proposition S7.*

PROPOSITION 1. *Suppose ψ is twice-differentiable and $\psi(0) = \psi'(0_+) = 0$. Suppose, for each $c > 0$, $u_c(v)$ is a sub-gamma or subexponential uniform boundary with crossing probability α for scale c. Then $v \mapsto u_{k_1}(k_2 v)$ is a sub-ψ uniform boundary for some constants $k_1, k_2 > 0$ depending only on ψ.*

Proposition 1 restates Howard et al. [25], Proposition 1, which shows that any process (S_t) which is sub-ψ is also sub-gamma and subexponential, if ψ satisfies the conditions of Proposition 1. Note that these conditions are satisfied for any mean-zero random variable if the CGF exists in a neighborhood of zero, so the conditions are quite weak [36], Theorem 2.3.

EXAMPLE 1 (Confidence sequence for the variance of a Gaussian distribution with unknown mean). Suppose $X_1, X_2, \ldots$ are i.i.d. draws from a $\mathcal{N}(\mu, \sigma^2)$ distribution and we wish to sequentially estimate σ^2 when μ is also unknown. Let $S_t := \sigma^{-2} \sum_{i=1}^{t+1} (X_i - \bar{X}_{t+1})^2 - t$ for $t = 1, 2, \ldots$, where $\bar{X}_t := t^{-1} \sum_{i=1}^{t} X_i$ is the sample mean. This S_t is a centered and scaled sample variance, and as in Darling and Robbins [12], we use the fact that S_t is a cumulative sum of independent, centered Chi-squared random variables each with one degree of freedom (see Appendix H for details). Such a centered Chi-squared distribution has variance two and CGF equal to $2\psi_{E,2}$.

Thus (S_t) is 1-subexponential with variance process $V_t = 2t$ and scale parameter $c = 2$. We may uniformly bound the upper deviations of S_t using any subexponential uniform boundary, for example, the gamma-exponential mixture boundary of Proposition S5. Or, we can use Proposition S7 to deduce that (S_t) is sub-gamma with scale $c = 2$ (and the same variance process) and use the closed-form stitched boundary of Theorem 1.

The above constructions yield lower confidence sequences for the variance. To obtain an upper confidence sequence, we use the fact that $(-S_t)$ is 1-subexponential with scale parameter $c = -2$. Now Proposition S7 implies that $(-S_t)$ is sub-gamma with scale $c = -1$, so the stitched boundary again applies, while Proposition S7 implies that $(-S_t)$ is also sub-Gaussian, so we may alternatively use the normal mixture boundary of Proposition S2. Since $\psi_{G,-1}$ is uniformly smaller than ψ_N, the above analysis yields tighter bounds than the sub-Gaussian approach of Darling and Robbins [12].

The simplest uniform boundaries are linear with positive intercept and slope. This is formalized in Howard et al. [25], partially restated below.

LEMMA 1 ([25], Theorem 1). *For any $\lambda \in [0, \lambda_{\max})$ and $\alpha \in (0, 1)$,*

$$u(v) := \frac{\log(l_0/\alpha)}{\lambda} + \frac{\psi(\lambda)}{\lambda} \cdot v \tag{2.5}$$

is a sub-ψ uniform boundary with crossing probability α.

While Lemma 1 provides a versatile building block, the $\mathcal{O}(V_t)$ growth of $u(V_t)$ may be undesirable. Indeed, from a concentration point of view, the typical deviations of S_t tend to be only $\mathcal{O}(\sqrt{V_t})$, so the bound will rapidly become loose for large t. From a confidence sequence point of view, recall that the confidence radius for the mean is given by $u(V_t)/t$. Typically, $V_t = \Theta(t)$ a.s. as $t \to \infty$, so the confidence radius will be asymptotically zero width if and only if $u(v) = o(v)$. In other words, we cannot achieve arbitrary estimation precision with arbitrarily large samples unless the uniform boundary is sublinear. We address this problem in Section 3, building upon Lemma 1 to construct *curved* sub-ψ uniform boundaries.

3. Curved uniform boundaries.

We present our four methods for computing curved uniform boundaries in Sections 3.1 to 3.4. In Section 3.5, we discuss how to tune boundaries, a necessity for good performance in practice, and we describe the unimprovability of sub-Gaussian mixture bounds in Section 3.6.

3.1. *Closed-form boundaries via stitching.*

Our analytical "stitched" bound is useful in the sub-Gaussian case or, more generally, the sub-gamma case with scale c. We require three user-chosen parameters:

- a scalar $\eta > 1$ determines the geometric spacing of intrinsic time,
- a scalar $m > 0$ which gives the intrinsic time at which the uniform boundary starts to be nontrivial, and
- an increasing function $h : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ such that $\sum_{k=0}^{\infty} 1/h(k) \leq 1$, which determines the shape of the boundary's growth after time m.

Recalling the scale parameter c for the ψ_G function above and the constant l_0 in Definition 1, we define the stitching function $\mathcal{S}_\alpha$ as

$$\mathcal{S}_\alpha(v) := \sqrt{k_1^2 v \ell(v) + k_2^2 c^2 \ell^2(v)} + k_2 c \ell(v),$$

$$\text{where} \begin{cases} \ell(v) := \log h\left(\log_\eta\left(\dfrac{v}{m}\right)\right) + \log\left(\dfrac{l_0}{\alpha}\right), \\ k_1 := (\eta^{1/4} + \eta^{-1/4})/\sqrt{2}, \\ k_2 := (\sqrt{\eta} + 1)/2, \end{cases} \tag{3.1}$$

and define the stitched boundary as $u(v) = \mathcal{S}_\alpha(v \vee m)$. Note $\mathcal{S}_\alpha(v) \leq k_1\sqrt{v\ell(v)} + 2ck_2\ell(v)$ when $c > 0$, while $\mathcal{S}_\alpha(v) \leq k_1\sqrt{v\ell(v)}$ when $c \leq 0$, with equality in the sub-Gaussian case ($c = 0$). These simpler expressions may sometimes be preferable. For notational simplicity, we suppress the dependence of $\mathcal{S}_\alpha$ on h, η, l_0 and c; we will discuss specific choices as necessary. In our examples, $\ell(v)$ grows as $\mathcal{O}(\log v)$ or $\mathcal{O}(\log\log v)$ as $v \uparrow \infty$, so the first term, $k_1\sqrt{V_t\ell(V_t)}$, dominates for sufficiently large V_t, specifically when $V_t/\ell(V_t) \gg 2c^2\sqrt{\eta}$.

THEOREM 1 (Stitched boundary). *For any $c \geq 0, \alpha \in (0, 1), \eta > 1, m > 0$ and $h : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ increasing such that $\sum_{k=0}^{\infty} 1/h(k) \leq 1$, the function $v \mapsto \mathcal{S}_\alpha(v \vee m)$ is a sub-gamma uniform boundary with crossing probability α. Further, for any sub-ψ_G process (S_t) with variance process (V_t) and any $v_0 \geq m$,*

$$\mathbb{P}(\exists t \geq 1 : V_t \geq v_0 \text{ and } S_t \geq \mathcal{S}_\alpha(V_t)) \leq \sum_{k=\lfloor \log_\eta(v_0/m) \rfloor}^{\infty} \frac{1}{h(k)}. \tag{3.2}$$

The first sentence above says that the probability of S_t crossing $\mathcal{S}_\alpha(V_t \vee m)$ at least once is at most α, while the second says that, even if it does happen to cross once or more, the probability of further crossings decays to zero beyond larger and larger intrinsic times v_0. Note that

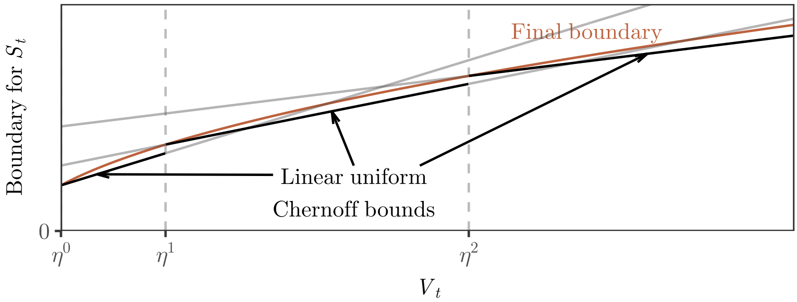

FIG. 3. *Illustration of Theorem 1, stitching together linear boundaries to construct a curved boundary. We break time into geometrically-spaced epochs $\eta^k \le V_t < \eta^{k+1}$, construct a linear uniform bound using Lemma 1 optimized for each epoch, and take a union bound over all crossing events. The final boundary is a smooth analytical upper bound to the piecewise linear bound.*

(3.2) implies $\mathbb{P}(\sup_t V_t = \infty \text{ and } S_t \ge \mathcal{S}_\alpha(V_t) \text{ infinitely often}) = 0$. The proof of Theorem 1, given with discussion in Appendix A.1, follows by taking a union bound over a carefully chosen family of linear boundaries, one for each of a sequence of geometrically-spaced epochs; see Figure 3. The high-level proof technique is standard, often referred to as "peeling" in the bandit literature, and closely related to chaining elsewhere in probability theory. Our proof generalizes beyond the sub-Gaussian case and involves careful parameter choices in order to achieve tight constants. In brief, within each epoch, there are many possible linear boundaries, and we have found that optimizing the linear boundary for the geometric mean of the epoch endpoints strikes a good balance between tight constants and analytical simplicity in the final boundary. Appendix G gives a detailed comparison of constants arising from our bound with similar bounds from the literature.

The boundary shape is determined by choosing the function h and setting the nominal crossing probability in the kth epoch to equal $\alpha/h(k)$. Then Theorem 1 gives a curved boundary which grows at a rate $\mathcal{O}(\sqrt{V_t \log h(\log_\eta V_t)})$ as $V_t \uparrow \infty$. The more slowly $h(k)$ grows as $k \uparrow \infty$, the more slowly the resulting boundary will grow as $V_t \uparrow \infty$. A simple choice is exponential growth, $h(k) = \eta^{sk}/(1-\eta^{-s})$ for some $s > 1$, yielding $\mathcal{S}_\alpha(v) = \mathcal{O}(\sqrt{v \log v})$. A more interesting example is $h(k) = (k+1)^s \zeta(s)$ for some $s > 1$, where $\zeta(s)$ is the Riemann zeta function. Then, when $l_0 = 1$, Theorem 1 yields the *polynomial stitched boundary*: for $c \ge 0$,

$$\begin{aligned} \mathcal{S}_\alpha(v) = {} & k_1 \sqrt{v\left(s \log\log\left(\frac{\eta v}{m}\right) + \log \frac{\zeta(s)}{\alpha \log^s \eta}\right)} \\ & + c k_2 \left(s \log\log\left(\frac{\eta v}{m}\right) + \log \frac{\zeta(s)}{\alpha \log^s \eta}\right), \end{aligned} \tag{3.3}$$

where the second term is neglected in the sub-Gaussian case since $c = 0$. This is a "finite LIL bound," so-called because $\mathcal{S}_\alpha(v) \sim \sqrt{s k_1^2 v \log\log v}$, matching the form of the law of the iterated logarithm [68]. We can bring $s k_1^2$ arbitrarily close to 2 by choosing η and s sufficiently close to one, at the cost of inflating the additive term $\log(\zeta(s)/(\log^s \eta))$. Briefly, increasing η increases the size of each epoch in the aforementioned peeling argument, which reduces the looseness of the union bound over epochs. But the larger we make the epochs, the further each linear boundary deviates from the ideal curved shape at the ends of the epochs, which inflates our final boundary. The choice of s involves a similar tradeoff: increasing s causes us to exhaust more of our total error probability budget on earlier epochs, decreasing the constant term (which matters most for early times), at the cost of a union bound over smaller error probabilities in later epochs, which shows up as an increase in the leading constant. We discuss parameter tuning in more practical terms in Section 3.5. For example, take

$\eta = 2, s = 1.4, m = 1$; if S_t is a sum of independent, zero-mean, 1-sub-Gaussian observations, we obtain

$$\mathbb{P}\left(\exists t \geq 1 : S_t \geq 1.7\sqrt{t\left(\log\log(2t) + 0.72\log\left(\frac{5.2}{\alpha}\right)\right)}\right) \leq \alpha. \tag{3.4}$$

Figure S2 in Appendix G compares a sub-Gaussian stitched boundary to a numerically-computed discrete mixture bound with a mixture distribution roughly corresponding to $h(k) \propto (k+1)^{1.4}$, as described in Appendix A.6. This discrete mixture boundary acts as a lower bound (see Section 3.6) and shows that not too much is lost by the approximations involved in the stitching construction. Figure S3 compare the same stitched boundary to related bounds from the literature; our bound shows slightly improved constants over the best known bounds.

Although our stitching construction begins with a sub-gamma assumption, it applies to other sub-ψ cases, including sub-Bernoulli, sub-Poisson and subexponential cases; see Figure 2 and Proposition 1. Further, our stitched bounds apply equally well in continuous-time settings to Brownian motion, continuous martingales, martingales with bounded jumps and martingales whose jumps satisfy a Bernstein condition; see Corollary S2.

While our focus is on nonasymptotic results, Theorem 1 makes it easy to obtain the following general upper asymptotic LIL, proved in Appendix A.2.

COROLLARY 1. *Suppose* (S_t) *is sub-*ψ *with variance process* (V_t) *and* $\psi(\lambda) \sim \lambda^2/2$ *as* $\lambda \downarrow 0$. *Then*

$$\limsup_{t\to\infty} \frac{S_t}{\sqrt{2V_t \log\log V_t}} \leq 1 \quad \text{on } \left\{\sup_t V_t = \infty\right\}. \tag{3.5}$$

3.2. *Conjugate mixture boundaries.* For appropriate choice of mixing distribution F, the integral $\int \exp\{\lambda S_t - \psi(\lambda)V_t\}\,\mathrm{d}F(\lambda)$ will be analytically tractable. Since, under Definition 1, this mixture process is upper bounded by a mixture supermartingale $\int L_t(\lambda)\,\mathrm{d}F(\lambda)$, such mixtures yield closed form or efficiently computable curved boundaries, which we call conjugate mixture boundaries. This approach is known as the method of mixtures, one of the most widely-studied techniques for constructing uniform bounds [14, 38, 41, 55, 57, 58, 72, 73]. Unlike the stitched bound of Theorem 1, which involves a small amount of looseness in the analytical approximations, mixture boundaries involve no such approximations and, in the sub-Gaussian case, are unimprovable in the sense described in Section 3.6. We restate the following standard idea behind the method of mixtures using our definitions, with a proof in Appendix A.3. The proof details a technical condition on product measurability which we require of L_t.

LEMMA 2. *For any probability distribution* F *on* $[0, \lambda_{\max})$ *and* $\alpha \in (0, 1)$,

$$\mathcal{M}_\alpha(v) := \sup\left\{s \in \mathbb{R} : \underbrace{\int \exp\{\lambda s - \psi(\lambda)v\}\,\mathrm{d}F(\lambda)}_{=:m(s,v)} < \frac{l_0}{\alpha}\right\} \tag{3.6}$$

is a sub-ψ *uniform boundary with crossing probability* α, *so long as the supermartingale* (L_t) *of Definition* 1 *is product measurable when the underlying probability space is augmented with the independent random variable* λ.

For each of our conjugate mixture bounds, we compute $m(s, v)$ in closed form. The boundary $u(v)$ can then be computed by numerically solving the equation $m(s, v) = l_0/\alpha$ in s, as we show in Appendix D. When an identical sub-ψ condition applies to $(-S_t)$ as well as (S_t), we may apply a uniform boundary to both tails and take a union bound, obtaining a two-sided confidence sequence. However, mixing over $\lambda \in \mathbb{R}$ rather than $\lambda \in \mathbb{R}_{\geq 0}$ yields a two-sided bound directly, so in some cases we present two-sided variants along with their one-sided counterparts. We give details for the following conjugate mixture boundaries in Appendix A.3:

- one-, two-sided *normal mixture* boundaries (sub-Gaussian case);
- one-, two-sided *beta-binomial mixture* boundaries (sub-Bernoulli case);
- one-sided *gamma-Poisson mixture* boundary (sub-Poisson case); and
- one-sided *gamma-exponential mixture* boundary (subexponential case).

The two-sided normal mixture boundary has a closed-form expression:

$$u(v) := \sqrt{(v+\rho)\log\left(\frac{l_0^2(v+\rho)}{\alpha^2\rho}\right)}. \tag{3.7}$$

The one-sided normal mixture boundary has a similar, closed-form upper bound, making these especially convenient. It is clear from (3.7) that the normal mixture boundary grows as $\mathcal{O}(\sqrt{v\log v})$ asymptotically, and this rate is shared by all of our conjugate mixture boundaries. Indeed, Proposition 2 below, proved in Appendix A.4, shows that such a rate holds for any mixture boundary as given by (3.6) whenever the mixing distribution is continuous with positive density at and around the origin, a property which holds for all mixture distributions used in our conjugate mixture boundaries, subject to regularity conditions on ψ which hold for the CGF of any nontrivial, mean-zero r.v. and specifically for the five ψ functions in Section 2.

PROPOSITION 2. *Assume* (*i*) *ψ is nondecreasing*, $\psi(0) = \psi'(0_+) = 0$, $\psi''(0_+) = c > 0$, *and ψ has three continuous derivatives on a neighborhood including the origin*; *and* (*ii*) *F has density f* (*w.r.t. Lebesgue*) *which is continuous and positive on a neighborhood including the origin. Then*

$$\mathcal{M}_\alpha(v) = \sqrt{v\left[c\log\left(\frac{cl_0^2 v}{2\pi\alpha^2 f^2(0)}\right) + o(1)\right]} \quad \text{as } v \to \infty. \tag{3.8}$$

Note that f need not place mass on all of $[0, \lambda_{\max})$, only near the origin, for the asymptotic rate to hold. Proposition 2 shows how the asymptotic behavior of any such mixture bound depends only on the behavior of ψ and f near the origin, a result reminiscent of the central limit theorem. Analogous, related results for the sub-Gaussian special case using $\psi(\lambda) = \lambda^2/2$ can be found in Robbins and Siegmund [58], Section 4, and Lai [42], Theorem 2, in some cases under weaker assumptions on F.

In contrast to previous derivations of conjugate mixture boundaries in the literature, all of our conjugate mixture boundaries include a common tuning parameter $\rho > 0$ which controls the sample size for which the boundary is optimized. Such tuning is critical in practice, as we explain in Section 3.5, but has been ignored in much prior work. Additionally, with the exception of the sub-Gaussian case, most prior work on the method of mixtures has focused on parametric settings. We instead emphasize the applicability of these bounds to nonparametric settings. For example, when the observations are bounded, one may construct a confidence sequence making use of empirical-Bernstein estimates (Theorem 4) based on our gamma-exponential mixture (Proposition S5). See Appendix J for other conditions in which mixture bounds yield nonparametric uniform boundaries.

3.3. *Numerical bounds using discrete mixtures.* In applications, one may not need an explicit closed-form expression so long as the bound can be easily computed numerically. Our discrete mixture method is an efficient technique for numerical computation of curved boundaries for processes satisfying Definition 1. It permits arbitrary mixture densities, thus producing boundaries growing at the rate $\mathcal{O}(\sqrt{v \log\log v})$. Recall that the shape of the stitched bound was determined by the user-specified function h. For the discrete mixture bound, one instead specifies a probability density f over finite support $(0, \bar{\lambda}]$ for some $\bar{\lambda} \in (0, \lambda_{\max})$. We first discretize f using a series of support points λ_k, geometrically spaced according to successive powers of some $\eta > 1$, and an associated set of weights w_k:

$$\lambda_k := \frac{\bar{\lambda}}{\eta^{k+1/2}} \quad \text{and} \quad w_k := \frac{\bar{\lambda}(\eta - 1) f(\lambda_k \sqrt{\eta})}{\eta^{k+1}} \quad \text{for } k = 0, 1, 2, \ldots. \tag{3.9}$$

THEOREM 2 (Discrete mixture bound). *Fix $\psi : [0, \lambda_{\max}) \to \mathbb{R}$, $\alpha \in (0, 1)$, $\bar{\lambda} \in (0, \lambda_{\max})$, and a probability density f on $(0, \bar{\lambda}]$ that is nonincreasing and positive. For supports λ_k and weights w_k defined in* (3.9),

$$\mathrm{DM}_\alpha(v) := \sup\left\{ s \in \mathbb{R} : \sum_{k=0}^{\infty} w_k \exp\{\lambda_k s - \psi(\lambda_k) v\} < \frac{l_0}{\alpha} \right\}, \tag{3.10}$$

is a sub-ψ uniform boundary with crossing probability α.

We suppress the dependence of DM_α on f, l_0, $\bar{\lambda}$ and η for notational simplicity. Though Theorem 2 is a straightforward consequence of the method of mixtures, our choice of discretization (3.9) makes it effective, broadly applicable and easy to implement. See Appendix A.5 for the proof of this result. Figure S2 includes an example bound, demonstrating a slight advantage over stitching. Appendix A.6 describes a connection between the stitching and discrete mixture methods, including a correspondence between the alpha-spending function h and the mixture density f. Finally, we note that the method can be applied even when f is not monotone; one must simply choose the discretization (3.9) more carefully, using known properties of f.

3.4. *Inverted stitching for arbitrary boundaries.* In the method of mixtures, we choose a mixing distribution F and the machinery yields a boundary $\mathcal{M}_\alpha$. Likewise, in the stitching construction of Theorem 1, we choose an error decay function h and obtain a boundary $\mathcal{S}_\alpha$. Here, we invert the procedure: we choose a boundary function $g(v)$ and numerically compute an upper bound on its S_t-upcrossing probability using a stitching-like construction.

THEOREM 3. *For any nonnegative, strictly concave function $g : \mathbb{R} \to \mathbb{R}$ and $v_{\max} > 1$, the function*

$$u(v) := \begin{cases} g(1 \vee v), & v \leq v_{\max}, \\ \infty, & \textit{otherwise} \end{cases} \tag{3.11}$$

is a sub-Gaussian uniform boundary with crossing probability at most

$$l_0 \inf_{\eta > 1} \sum_{k=0}^{\lceil \log_\eta v_{\max} \rceil} \exp\left\{ -\frac{2(g(\eta^{k+1}) - g(\eta^k))(\eta g(\eta^k) - g(\eta^{k+1}))}{\eta^k (\eta - 1)^2} \right\}. \tag{3.12}$$

The proof is in Appendix A.7. For simplicity, we restrict to the sub-Gaussian case; examination of the proof will show that the method applies in other sub-ψ cases as well, since we simply apply Lemma 1 to appropriately chosen lines, but more involved numerical calculations will be necessary, as the closed form (3.12) no longer applies. A similar idea was considered by Darling and Robbins [14], using a mixture integral approximation instead of an epoch-based construction to derive closed-form bounds. Theorem 3 requires numerical summation but yields tighter bounds with fewer assumptions. As an example, Theorem 3 with $\eta = 2.99$ shows that

$$\mathbb{P}\left(\exists t : 1 \leq V_t \leq 10^{20} \text{ and } S_t \geq 1.7\sqrt{V_t\left(\log\log(eV_t) + 3.46\right)}\right) \leq 0.025. \tag{3.13}$$

This boundary is illustrated in Figure S2.

3.5. *Tuning boundaries in practice.* All uniform boundaries involve a tradeoff of tightness at different intrinsic times: making a bound tighter for some range of times requires making it looser at other times. Roughly speaking, the choice of a uniform boundary involves choosing both what time the bound should be optimized for (e.g., should the bound be tightest around 100 observations or around 100,000 observations?) as well as how quickly the bound degrades as we move away from the optimized-for time (e.g., if we optimize for 100 samples, will the bound be twice as wide when we reach 1000 samples, or will it stay within a factor of two until we reach 1,000,000 samples?). A boundary which decays more slowly will necessarily not be as tight around the optimized-for time. In brief, linear boundaries decay the most quickly, conjugate mixture boundaries decay substantially more slowly, and polynomial stitched boundaries decay even more slowly; we feel that mixture boundaries strike a good balance in practice.

Here, we explain how to optimize uniform boundaries for a particular time and discuss the above tradeoff in more detail. Let $W_{-1}(x)$ be the lower branch of the Lambert W function, the most negative real-valued solution in z to $ze^z = x$. Consider the unitless process $S_t/\sqrt{V_t}$, and the corresponding uniform boundary $v \mapsto u(v)/\sqrt{v}$. Since all of our uniform boundaries $u(v)$ have positive intercept at $v = 0$, and all grow at least at the rate $\sqrt{v \log\log v}$ as $v \to \infty$, the normalized boundary $u(v)/\sqrt{v}$ diverges as $v \to 0$ and $v \to \infty$. For the two-sided normal mixture (3.7), there is a unique time m at which $u(v)/\sqrt{v}$ is minimized; m is proportional to tuning parameter ρ as follows:

PROPOSITION 3. *Let $u(v)$ be the two-sided normal mixture boundary* (3.7) *with parameter $\rho > 0$.*

(a) *For fixed $\rho > 0$, the function $v \mapsto u(v)/\sqrt{v}$ is uniquely minimized at $v = m$ with m given by*

$$\frac{m}{\rho} = -W_{-1}\left(-\frac{\alpha^2}{el_0^2}\right) - 1. \tag{3.14}$$

(b) *For fixed $m > 0$, the choice of ρ which minimizes the boundary value $u(m)$ is also determined by* (3.14).

The above result is proved in Appendix C.1; it is a matter of elementary calculus, but addresses a question that has received little attention in the literature. Figure 4 includes the normalized versions of two normal mixture boundaries optimized for different times, $m =$ 300 and $m = 5000$. Optimizing for the range of values of V_t most relevant in a particular application will yield the tightest confidence sequences. However, as the Figure shows, one need not have a very precise range of times, so long as one uses a conservatively low value for m, because $u(v)/\sqrt{v}$ grows slowly after time m. Indeed, for the normal mixture boundary with $\alpha = 0.05$ and $l_0 = 1$, we have $u(m)/\sqrt{m} \approx 3.0$ and $u(100m)/\sqrt{100m} \approx 3.6$, so that the penalty for being off by two orders of magnitude is modest.

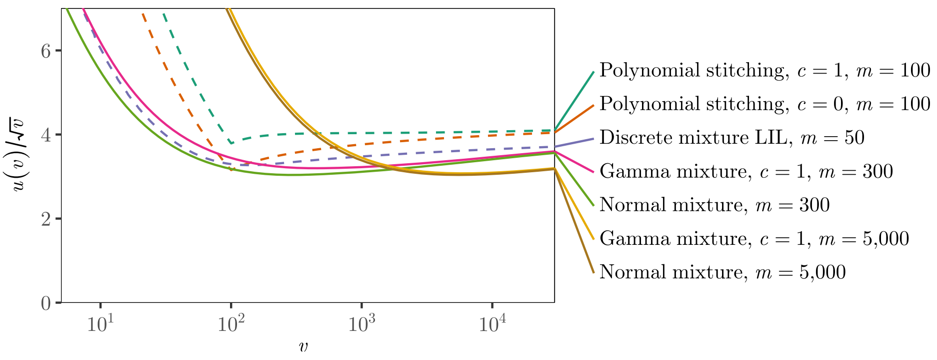

FIG. 4. *Comparison of normalized uniform boundaries $u(v)/\sqrt{v}$ optimized for different intrinsic times. Normal mixture uses Appendix Proposition S2, while gamma mixture uses Appendix Proposition S5. Polynomial stitched boundary is given in* (3.3), *with $\eta = 2$ and $s = 1.4$. Discrete mixture applies Theorem* 2 *to the density $f(\lambda) = 0.4 \cdot 1_{0 \le \lambda \le 0.38}/[\lambda \log^{1.4}(0.38e/\lambda)]$ with $\eta = 1.1$, and $\lambda_{\max} = 0.38$; see Appendix A.6 for motivation. All boundaries use $\alpha = 0.025$.*

The one-sided normal mixture boundary of Appendix Proposition S2 with crossing probability α is nearly identical to the two-sided normal mixture boundary with crossing probability 2α, so one may choose ρ as in Proposition 3 with α doubled. For the gamma-exponential mixture and other non-sub-Gaussian uniform boundaries, Proposition 3 provides a good approximation in practice. Figure 4 includes gamma-exponential mixture boundaries with the same ρ values as each corresponding normal mixture boundary. Though the normalized gamma-exponential mixture boundary with $m = 300$ clearly reaches its minimum at $v > m$, this choice of ρ seems reasonable. Discrete mixtures can be similarly tuned by adjusting the precision of the mixing distribution, but require additional considerations (Appendix E).

Comparing the sub-Gaussian stitched boundary, discrete mixture boundary and normal mixture boundary optimized for $m = 300$ in Figure 4 illustrates another important point for practice: although the normal mixture bound grows more quickly than the others as $v \to \infty$, it remains smaller over about three orders of magnitude. This makes it preferable for many real-world applications, as the longest feasible duration of an experiment is rarely more than two orders of magnitude larger than the earliest possible stopping time. For example, many online experiments run for at least one week to account for weekly seasonality effects, and very few such experiments last longer than 100 weeks. As both the normal mixture and the discrete mixture are unimprovable in general (Section 3.6), the difference is attributable to the choice of mixture, or alternatively, to the fact that the normal mixture trades tightness around the optimized-for time in exchange for looseness at much later times. The lesson is that the $\mathcal{O}(v \log\log v)$ rate, while asymptotically optimal in certain settings and useful for theory and some applications, may not be preferable in all real-world scenarios.

3.6. *Unimprovability of uniform boundaries.* Definition 2 of a sub-ψ boundary u involves only an upper bound on the u-crossing probability of any sub-ψ process (S_t). One may reasonably ask for corresponding lower bounds on the u-crossing probability to quantify how tight this boundary is. In the ideal case, we might desire a boundary u such that the true u-crossing probability of some process (S_t) is equal to the upper bound. In nonparametric settings, we cannot achieve this goal for every sub-ψ process. However, we might still ask that there exists *some* sub-ψ process for which the true u-crossing probability is arbitrarily close to the upper bound, so that the latter is unimprovable in general. That is, we might ask that the inequality on the supremum in Definition 2 holds with equality.

The fact we wish to point out, known in various forms, is that in the scalar, sub-Gaussian case, exact mixture bounds are unimprovable in the above sense. It is in this sense that the discrete mixture bound in Figure S2 provides a lower bound, showing that the sub-Gaussian polynomial stitched bound cannot be improved by much. The following result shows that, for any exact, sub-Gaussian mixture boundary $\mathcal{M}_\alpha$, as defined in Lemma 2 for $\psi = \psi_N$, there exists a sub-Gaussian process whose true $\mathcal{M}_\alpha$-crossing probability is arbitrarily close to α. The result is similar to Theorem 2 of Robbins and Siegmund [58], which gives a more general invariance principle, but requires conditions on the boundary that appear difficult to verify for arbitrary mixture boundaries $\mathcal{M}_\alpha$. Recall that $\mathbb{S}^1_{\psi_N}$ is the class of pairs of processes (S_t, V_t) such that (S_t) is 1-sub-Gaussian with variance process (V_t).

PROPOSITION 4. *For any exact*, 1-*sub-Gaussian mixture boundary* $\mathcal{M}_\alpha$,

$$\sup_{(S_t,V_t)\in\mathbb{S}^1_{\psi_N}} \mathbb{P}\big(\exists t \geq 1 : S_t \geq \mathcal{M}_\alpha(V_t)\big) = \alpha. \tag{3.15}$$

We prove Proposition 4 in Appendix C.2. In general, for each α there is an infinite variety of boundaries that are unimprovable in the above sense, differing in when they are loose and tight. These different boundaries will yield confidence sequences which are loose or tight at different sample sizes, or, equivalently, are efficient for detecting different effect sizes. Such a boundary cannot be tightened everywhere without increasing the crossing probability.

4. Applications.

After presenting an empirical-Bernstein confidence sequence for bounded observations, we apply our techniques to causal effect estimation and matrix martingales. We also consider estimation for a general, one-parameter exponential family.

4.1. *An empirical-Bernstein confidence sequence.*

The following novel result is proved in Appendix A.8 using a self-normalization argument, which leads to its attractive simplicity. Recall the estimand $\mu_t := t^{-1}\sum_{i=1}^t \mathbb{E}_{i-1}X_i$, the average conditional expectation.

THEOREM 4. *Suppose* $X_t \in [a, b]$ *a.s. for all* t. *Let* $(\widehat{X}_t)$ *be any* $[a, b]$-*valued predictable sequence*, *and let* u *be any subexponential uniform boundary with crossing probability* α *for scale* $c = b - a$. *Then*

$$\mathbb{P}\left(\forall t \geq 1 : |\bar{X}_t - \mu_t| < \frac{u(\sum_{i=1}^t (X_i - \widehat{X}_i)^2)}{t}\right) \geq 1 - 2\alpha. \tag{4.1}$$

This is an empirical-Bernstein bound because it uses the sum of observed squared deviations to estimate the true variance, much like a classical t-test. Hence the confidence radius scales with the true standard deviation for sufficiently large samples, regardless of the support diameter $b - a$, and with no prior knowledge of the true variance. Note also that this bound does not require that observations share a common mean.

The confidence statement (4.1) holds for *any* sequence of predictions $(\widehat{X}_i)$, but predictions closer to the conditional expectations, $\widehat{X}_i \approx \mathbb{E}_{i-1}X_i$, will yield smaller confidence intervals on average. A simple choice is the mean, $\widehat{X}_t = (t-1)^{-1}\sum_{i=1}^{t-1} X_i$, which will be effective when the samples are i.i.d., for example. But the predictions $(\widehat{X}_i)$ can also make use of trends, seasonality, stratification or regression (in the presence of covariates), machine learning algorithms or any other information that may aid with prediction.

For an explicit example, assume $X_i \in [0, 1]$ and define the empirical variance as $\widehat{V}_t := \sum_{i=1}^t (X_i - \bar{X}_{i-1})^2$. Invoking Theorem 4 with the boundary (3.3) using $c = 1$, $\eta = 2$, $m = 1$, and $h(k) \propto k^{1.4}$, we have the following 95% confidence sequence for μ_t:

$$\bar{X}_t \pm \frac{1.7\sqrt{(\widehat{V}_t \vee 1)(\log\log(2(\widehat{V}_t \vee 1)) + 3.8)} + 3.4\log\log(2(\widehat{V}_t \vee 1)) + 13}{t}. \tag{4.2}$$

When a closed form is not required, the gamma-exponential mixture (supplement Proposition S5, see [26]) may yield tighter bounds than stitching; simulations in Section 5 demonstrate the use of Theorem 4 with this mixture.

4.2. *Estimating ATE in the Neyman–Rubin model.* As one illustration of Theorem 4, we consider the sequential estimation of average treatment effect under the Neyman–Rubin potential outcomes model [27, 61, 67]. We imagine a sequence of experimental units, each with real-valued potential outcomes under control and treatment denoted by $\{Y_t(0), Y_t(1)\}_{t\in\mathbb{N}}$, respectively. These potential outcomes are fixed, but we observe only one outcome for each unit in the experiment. We assign a randomized treatment to each unit, denoted by the $\{0, 1\}$-valued random variable $Z_t \in \mathcal{F}_t$, observing $Y_t^{\text{obs}} := Y_t(Z_t)$. Here, treatment is assigned by flipping a coin for each subject, with a bias possibly depending on previous observations. This treatment assignment is the only source of randomness. Specifically, let $P_t := E_{t-1}Z_t$ and suppose $0 < P_t < 1$ a.s. for all t; then we permit P_t to vary between individuals and to depend on past outcomes. This accommodates Efron's biased coin design [19] and related covariate balancing methods.

At each step t, having treated and observed units $1, \dots, t$, we wish to draw inference about the estimand $\text{ATE}_t := t^{-1}\sum_{i=1}^t [Y_i(1) - Y_i(0)]$. In particular, we seek a confidence sequence for $(\text{ATE}_t)_{t=1}^\infty$. To construct our estimator, we may utilize any predictions $\widehat{Y}_t(0)$ and $\widehat{Y}_t(1)$ for each unit's potential outcomes; these random variables must be $\mathcal{F}_{t-1}$-measurable, for each t. We then employ the inverse probability weighting estimator

$$X_t := \widehat{Y}_t(1) - \widehat{Y}_t(0) + \left(\frac{Z_t - P_t}{P_t(1 - P_t)}\right)\left(Y_t^{\text{obs}} - \widehat{Y}_t(Z_t)\right), \tag{4.3}$$

which is (conditionally) unbiased for the individual treatment effect $Y_t(1) - Y_t(0)$. As with Theorem 4, better predictions will lead to shorter confidence intervals, but the coverage guarantee holds for any choice of predictions, and a reasonable choice would be the average of past observed outcomes. See Aronow and Middleton [2] for a similar strategy for fixed-sample estimation.

We assume bounded potential outcomes; for simplicity, we assume $Y_t(k) \in [0, 1]$ for all $t \geq 1$, $k = 0, 1$, and we assume predictions are likewise bounded. We further assume that treatment probabilities are uniformly bounded away from zero and one. Then an empirical-Bernstein confidence sequence for ATE_t follows from Theorem 4, where we use $\widehat{X}_t = \widehat{Y}_t(1) - \widehat{Y}_t(0)$ so that

$$V_t := \sum_{i=1}^t (X_i - \widehat{X}_i)^2 = \sum_{i=1}^t \left(\frac{Z_i - P_i}{P_i(1 - P_i)}\right)^2 \left(Y_i^{\text{obs}} - \widehat{Y}_i(Z_i)\right)^2. \tag{4.4}$$

COROLLARY 2. *Suppose $P_t \in [p_{\min}, 1 - p_{\min}]$ a.s., $Y_t(k) \in [0, 1]$ and $\widehat{Y}_t(k) \in [0, 1]$ for all $t \geq 1$, $k = 0, 1$. Let u be any subexponential uniform boundary with scale $2/p_{\min}$ and crossing probability α. Then*

$$\mathbb{P}\left(\forall t \geq 1 : |\bar{X}_t - \text{ATE}_t| < \frac{u(V_t)}{t}\right) \geq 1 - 2\alpha. \tag{4.5}$$

For u, one may choose the gamma-exponential mixture boundary (supplement Proposition S5) or the stitched boundary (3.3) with $c = \frac{2}{p_{\min}}$. Figure 5 illustrates our strategy on simulated data. Over the range $t = 100$ to $t = 100{,}000$ displayed, our bound is about twice as wide as the fixed-sample CLT bound, with the ratio growing at a slow $\mathcal{O}(\sqrt{\log t})$ rate thereafter. Of course, the fixed-sample CLT bound provides no uniform coverage guarantee.

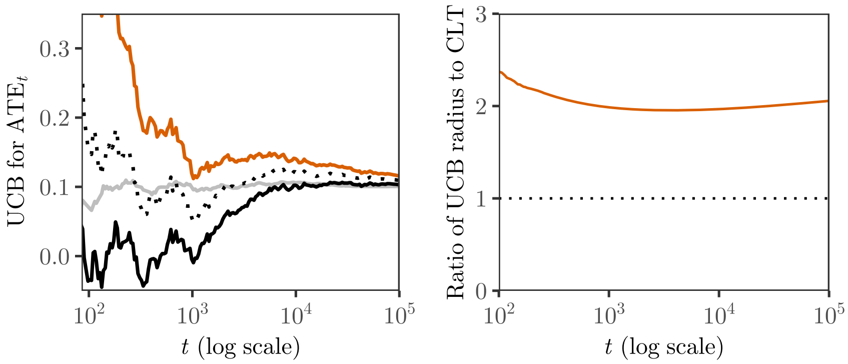

FIG. 5. *Upper half of* 95% *empirical-Bernstein confidence sequence for* ATE_t *under Bernoulli randomization based on one simulated sequence of i.i.d. observations*, $P_t \equiv 0.5$, $Y_i(0) \sim \text{Ber}(0.5)$, $Y_i(1) = \xi_i \vee Y_i(0)$ *where* $\xi_i \sim \text{Ber}(0.2)$. *Grey line shows estimand* ATE_t. *Dotted line shows fixed-sample confidence bounds based on difference-in-means estimator and normal approximation*; *these bounds fail to cover the true* ATE_t *at many times. Our bound uses* $\widehat{Y}_t(k) = \sum_{i=1}^{t-1} Y_i^{\text{obs}} 1_{Z_i = k} / \sum_{i=1}^{t-1} 1_{Z_i=k}$, $\alpha = 0.05$ *and a gamma-exponential mixture bound with* $\rho = 12.6$, *chosen to optimize for intrinsic time* $V_t = 100$.

4.3. *Matrix iterated logarithm bounds*. Our second application is the construction of iterated logarithm bounds for random matrix sums and their use in sequential covariance matrix estimation. The curved uniform bounds given in Section 3 may be applied to matrix martingales by taking (S_t) to be the maximum eigenvalue process of the martingale and (V_t) the maximum eigenvalue of the corresponding matrix variance process. Howard et al. [25], Section 2, give sufficient conditions for Definition 1 to hold in this matrix case. Then Theorem 1 yields a novel matrix finite LIL; here, we give an example for bounded increments. We denote the space of symmetric, real-valued, $d \times d$ matrices by $\mathbb{S}^d$; $\gamma_{\max}(\cdot)$ denotes the maximum eigenvalue; $\ell_{\eta,s}(v) = s \log\log(\eta v/m) + \log \frac{d\zeta(s)}{\alpha \log^s \eta}$; and $k_1(\eta)$, $k_2(\eta)$ are defined in (3.1).

COROLLARY 3. *Suppose* $(Y_t)_{t=1}^\infty$ *is a* $\mathbb{S}^d$*-valued matrix martingale such that* $\gamma_{\max}(Y_t - Y_{t-1}) \leq b$ *a.s. for all* t. *Let* $V_t := \gamma_{\max}(\sum_{i=1}^t \mathbb{E}_{t-1}(Y_t - Y_{t-1})^2)$ *and* $S_t := \gamma_{\max}(Y_t)$. *Then for any* $\eta > 1, s > 1, m > 0, \alpha \in (0,1)$, *we have*

$$\mathbb{P}\left(\exists t \geq 1 : S_t \geq k_1(\eta)\sqrt{(V_t \vee m)\ell_{\eta,s}(V_t \vee m)} + \frac{bk_2(\eta)}{3}\ell_{\eta,s}(V_t \vee m)\right) \leq \alpha. \tag{4.6}$$

The result follows using the polynomial stitched boundary after invoking Fact 1(c) and Lemma 2 of Howard et al. [25] (cf. [69]), which show that (S_t) is sub-gamma with variance process (V_t), scale $c = b/3$, and $l_0 = d$. Beyond bounded increments, the same bound holds for any sub-gamma process. As evidenced by Proposition 1, this is a very general condition.

Taking η and s arbitrarily close to one and using the final result of Theorem 1, we obtain the following asymptotic matrix upper LIL, proved in Appendix A.9. Here, we denote the martingale increments by $\Delta Y_t := Y_t - Y_{t-1}$.

COROLLARY 4. *Let* $(Y_t)_{t=1}^\infty$ *be a* $\mathbb{S}^d$*-valued, square-integrable martingale, and define* $V_t = \gamma_{\max}(\sum_{i=1}^t \mathbb{E}_{i-1}\Delta Y_t^2)$. *Then*

$$\limsup_{t\to\infty} \frac{\gamma_{\max}(Y_t)}{\sqrt{2V_t \log\log V_t}} \leq 1 \quad \textit{a.s. on } \left\{\sup_t V_t = \infty\right\} \tag{4.7}$$

whenever either (1) *the increments* (ΔY_t) *are i.i.d., or* (2) *the increments* (ΔY_t) *satisfy a Bernstein condition on higher moments*: *for some* $c > 0$, *for all* t *and all* $k > 2$, $\mathbb{E}_{t-1}(\Delta Y_t)^k \preceq (k!/2)c^{k-2}\mathbb{E}_{t-1}\Delta Y_t^2$.

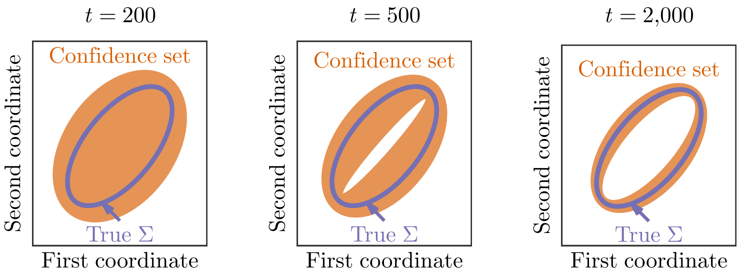

FIG. 6. *The matrix confidence sequence of Corollary 5 based on one simulated sequence. Observations are drawn i.i.d. taking values* $\pm(\sqrt{2}\ \sqrt{2})^T$, $\pm(1/\sqrt{2}\ -1/\sqrt{2})^T$ *each with probability* $1/4$, *with covariance matrix* $\Sigma = \frac{1}{4}\left(\begin{smallmatrix}5 & 3\\ 3 & 5\end{smallmatrix}\right)$, *which is represented by the ellipse* $x^T\Sigma^{-1}x = 1$. *Confidence ball with level* $\alpha = 0.05$ *is represented by shaded area between ellipses corresponding to elements of the confidence ball with minimal and maximal trace. Confidence sequence from Corollary 5 uses* $b = 4$ *and a discrete mixture boundary with* $\psi = \psi_G$ *using* $c = 2b/3$, *mixture density* $f_{1.4}^{\mathrm{LIL}}$ *from* (A.51) *with* $s = 1.4$ *matching* (3.4), $\eta = 1.1$ *and* $\overline{\lambda} = 0.262$ *chosen as described in Appendix E.*

The Bernstein condition holds if the increments are uniformly bounded, $\gamma_{\max}(\Delta Y_t) \le c$ for some $c > 0$. Also, in the i.i.d. case, $\mathbb{P}(V_t \to \infty) = 1$ and then (4.7) states that $\limsup_{t\to\infty} \gamma_{\max}(Y_t)/\sqrt{2\gamma_{\max}(\mathbb{E}\Delta Y_1^2)t\log\log t} \le 1$, a.s. on $\{\sup_t V_t = \infty\}$. When $d = 1$, this recovers the classical upper LIL, showing that Corollary 4 cannot be improved uniformly, but we are not aware of an appropriate lower bound for the general matrix case.

We now consider the nonasymptotic sequential estimation of a covariance matrix based on bounded vector observations [22, 39, 62, 70, 71]. In particular, we observe a sequence of independent, mean zero, $\mathbb{R}^d$-valued random vectors x_t with common covariance matrix $\Sigma = \mathbb{E}x_t x_t^T$. We wish to estimate Σ using an operator-norm confidence ball centered at the empirical covariance matrix $\widehat{\Sigma}_t := t^{-1}\sum_{i=1}^t x_i x_i^T$. For fixed-sample estimation, when $\|x_i\|_2 \le \sqrt{b}$ a.s. for all $i \in [t]$, the analysis of Tropp [70], Section 1.6.3, implies

$$\mathbb{P}\left(\|\widehat{\Sigma}_t - \Sigma\|_{\mathrm{op}} \ge \sqrt{\frac{2b\|\Sigma\|_{\mathrm{op}}\log(2d/\alpha)}{t}} + \frac{4b\log(2d/\alpha)}{3t}\right) \le \alpha. \tag{4.8}$$

We use a sub-Poisson uniform boundary to obtain a uniform analogue.

COROLLARY 5. *Let* $(x_t)_{t=1}^\infty$ *be a sequence of* $\mathbb{R}^d$*-valued, independent random vectors with* $\mathbb{E}x_t = 0$, $\|x_t\|_2 \le \sqrt{b}$ *a.s. and* $\mathbb{E}x_t x_t^T = \Sigma$ *for all* t. *If* u *is a sub-Poisson uniform boundary with crossing probability* α *and scale* $2b$, *then*

$$\mathbb{P}\left(\exists t \ge 1 : \|\widehat{\Sigma}_t - \Sigma\|_{\mathrm{op}} \ge \frac{1}{t}u\big(bt\|\Sigma\|_{\mathrm{op}}\big)\right) \le \alpha. \tag{4.9}$$

For example, using the polynomial stitched bound with scale $c = 2b/3$ and $m = b\|\Sigma\|_{\mathrm{op}}$, Corollary 5 gives a $(1-\alpha)$-confidence sequence for Σ with operator norm radius $\mathcal{O}(\sqrt{t^{-1}\log\log t})$. This bound has the closed form

$$\mathbb{P}\left(\exists t \ge 1 : \|\widehat{\Sigma}_t - \Sigma\|_{\mathrm{op}} \ge k_1\sqrt{\frac{b\|\Sigma\|_{\mathrm{op}}\ell(t)}{t}} + \frac{4bk_2\ell(t)}{3t}\right) \le \alpha, \tag{4.10}$$

where $\ell(t) = s\log\log(\eta t) + \log\frac{d\zeta(s)}{\alpha\log^s\eta}$, and k_1, k_2 are defined in (3.1).

In other words, with high probability, we have for all $t \ge 1$ that

$$\|\widehat{\Sigma}_t - \Sigma\|_{\mathrm{op}} \lesssim \sqrt{\frac{b\log(d\log t)}{t}} + \frac{b\log(d\log t)}{t}. \tag{4.11}$$

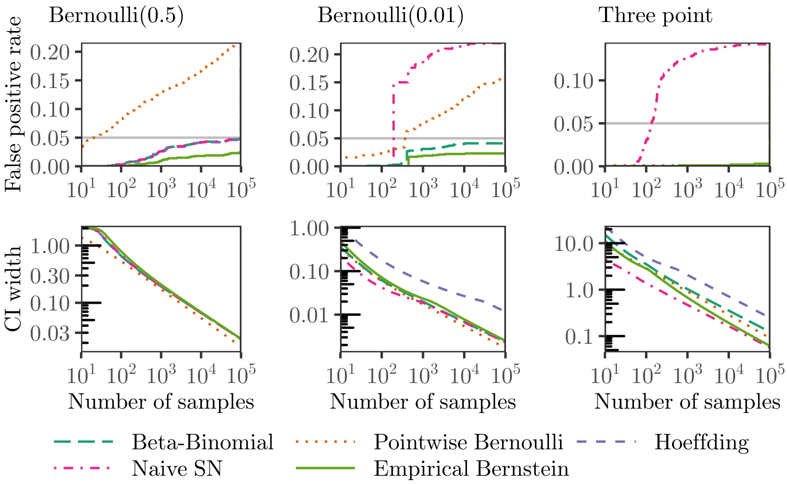

FIG. 7. *Summary of* 1000 *simulations, each with* 100,000 *i.i.d. observations from the indicated distribution. Top panels show the proportion of replications in which the* 95%*-confidence sequence has excluded the true mean by time* t. *Bottom panels show the mean confidence interval width. The "three point" distribution takes values* -1.408 *and* 1 *with probability* 0.495 *each, and takes value* 20 *with probability* 0.01. *"Hoeffding" uses a normal mixture boundary* (3.7), *while"Beta-Binomial" uses the beta-binomial mixture* (*Proposition S*3). *"Pointwise Bernoulli" uses a nonasymptotic bound based on the Bernoulli KL-divergence, which is valid pointwise but not uniformly. "Empirical Bernstein" uses the strategy given in Theorem* 4 *with a gamma-exponential mixture boundary, Proposition S*5. *"Naive SN" uses a normal mixture boundary with an empirical variance estimate, which does not guarantee coverage. In all cases,* ρ *is chosen to optimize for a sample size of* $t = 500$.

Compared to the fixed-sample result (4.8), we obtain uniform control by adding a factor of $\log\log t$. We are not aware of other results like these for sequential covariance matrix estimation. Figure 6 illustrates the confidence sequence of Corollary 5 on simulated data using a discrete mixture boundary with the mixture density f_s^{LIL} defined in (A.51).

4.4. *One-parameter exponential families.* Suppose (X_t) are i.i.d. from an exponential family in mean parametrization, with sufficient statistic $T(X)$ having mean in some set Ω. For each $\mu \in \Omega$, we write the density as $f_\mu(x) = h(x)\exp\{\theta(\mu)T(x) - A(\theta(\mu))\}$ where $A'(\theta(\mu)) = \mu$. Let ψ_μ be the cumulant-generating function of $T(X_1) - \mu$ when $\mathbb{E}T(X_1) = \mu$, that is, $\psi_\mu(\lambda) := A(\lambda + \theta(\mu)) - A(\theta(\mu)) - \lambda\mu$, with $\psi_\mu(\lambda) := \infty$ if the RHS does not exist. Writing $S_t(\mu) := \sum_{i=1}^t T(X_i) - t\mu$, the process $\exp\{\lambda S_t(\mu) - t\psi_\mu(\lambda)\}$ is the likelihood ratio testing $H_0 : \theta = \theta(\mu)$ against $H_1 : \theta = \theta(\mu) + \lambda$, and if we use a method-of-mixtures uniform boundary, the resulting confidence sequence will be dual to a family of mixture sequential probability ratio tests, as discussed in Section 6. To obtain a two-sided confidence sequence, we use the "reversed" CGF $\tilde{\psi}_\mu(\lambda) = \psi_\mu(-\lambda)$. We summarize these observations as follows; see Lai [41], Theorem 1, for a related result.

COROLLARY 6. *Suppose, for each* $\mu \in \Omega$, u_μ *is a sub-*ψ_μ *uniform bound with crossing probability* α_1, *and* $\tilde{u}_\mu$ *is a sub-*$\tilde{\psi}_\mu$ *uniform bound with crossing probability* α_2. *Defining*

$$\mathrm{CI}_t := \{\mu \in \Omega : -\tilde{u}_\mu(t) < S_t(\mu) < u_\mu(t)\}, \tag{4.12}$$

we have $\mathbb{P}(\forall t \geq 1 : \mathbb{E}T(X_1) \in \mathrm{CI}_t) \geq 1 - \alpha_1 - \alpha_2$.

5. Simulations. In[1] Figure 7, we illustrate the error control of some of our confidence sequences for estimating the mean of an i.i.d. sequence of observations (X_i) with bounded support $[a, b]$. We compare four strategies:

1. The Hoeffding strategy exploits the fact that bounded observations are sub-Gaussian ([24]; cf. [25], Lemma 3(c)). We use a two-sided normal mixture boundary (3.7) with variance process $V_t = (b-a)^2 t/4$.

2. The beta-binomial strategy uses the stronger condition that bounded observations are sub-Bernoulli ([24]; cf. [25], Fact 1(b)), accounting for the true mean as well as the boundedness, but possibly failing to take account of the true variance. For hypothesized true mean μ, this strategy uses the beta-binomial mixture boundary given in Proposition S3, with parameters $g(\mu) = \mu - a$ and $h(\mu) = b - \mu$, and variance process $V_t(\mu) = g(\mu)h(\mu)t$. The confidence set for the mean is $\{\mu \in [a, b] : -f_{g(\mu),h(\mu)}(V_t(\mu)) \le \sum_{i=1}^t X_i - t\mu \le f_{h(\mu),g(\mu)}(V_t(mu))\}$. This is more efficiently computed using the mixture supermartingale $m(S_t, V_t)$ of (A.23), as $\{\mu \in [a, b] : m(\sum_{i=1}^t X_i - t\mu, V_t(\mu)) < 1/\alpha\}$.

3. The pointwise Bernoulli strategy uses the same sub-Bernoulli condition as the beta-binomial strategy, but relies on a fixed-sample Cramér–Chernoff bound which is valid pointwise but not uniformly over time. Specifically, we reject mean μ if $V_t \psi_B^\star(S_t/V_t) \ge \log \alpha^{-1}$, where S_t is the sum of centered observations as usual, $V_t = (\mu - a)(b - \mu)t$, and we set $g = \mu - a$, $h = b - \mu$ in ψ_B, with $\psi_B^\star$ its Legendre–Fenchel transform.

4. The empirical-Bernstein strategy uses an empirical estimate of variance, thus achieving a confidence width scaling with the true variance in all three cases. Here, we use Theorem 4 with a gamma-exponential mixture boundary (supplement Proposition S5). For predictions, we use the mean of past observations: $\widehat{X}_t = (t-1)^{-1} \sum_{i=1}^{t-1} X_i$.

5. The naive self-normalized ("Naive SN") strategy plugs the empirical variance estimate, the sum of squared prediction errors from Theorem 4, into the two-sided normal mixture (3.7). It ignores the facts that the observations are not sub-Gaussian with respect to their true variance and that the variance is estimated. This strategy is similar to that of Johari et al. [34] and does not guarantee coverage. Though it will sometimes control false positives, coverage rates can easily be inflated for asymmetric, heavy-tailed distributions, as we illustrate.

We present three cases of bounded distributions. The first case is the easiest, with Ber(0.5) observations. Here, the sub-Gaussian variance parameter based on the boundedness of the observations is equal to the true variance, so the Hoeffding strategy performs well. The empirical-Bernstein strategy is only a little wider, and all four successfully control false positives. The story changes with the more difficult Ber(0.01) distribution, however. The Hoeffding boundary is far too wide, since it fails to make use of information about the true variance. The beta-binomial bound uses information about variance provided by the first moment to achieve the correct scaling. The naive self-normalized strategy, on the other hand, yields confidence intervals that are too small and fail to control false positive rate. The empirical Bernstein strategy, though only slightly wider than the naive bound for large sample sizes, gives just enough extra width to control the false positive rate and is nearly as narrow as the beta-binomial bound. The final, three-point distribution takes values -1.408 and 1 with probability 0.495 each, and takes value 20 with probability 0.01. Here, the beta-binomial strategy yields confidence intervals that are too wide. In this most difficult case, only the empirical Bernstein strategy yields tight intervals while controlling false positive rates.

[1] The repository https://github.com/gostevehoward/cspaper contains code to reproduce all simulations and plots in this paper. Uniform boundaries themselves are implemented in R and Python packages at https://github.com/gostevehoward/confseq.

6. Implications for sequential hypothesis testing. We have organized our presentation around confidence sequences and closely related uniform concentration bounds due to our belief that they offer a useful "user interface" for sequential inference. However, our methods also yield always-valid p-values [35] for sequential tests. Indeed, a slew of related definitions from the literature are equivalent or "dual" to one another. Here, we briefly discuss these connections. The following result, proved in Appendix C.4, gives equivalent formulations of common definitions in sequential testing.

LEMMA 3. *Let $(A_t)_{t=1}^{\infty}$ be an adapted sequence of events in some filtered probability space and let $A_\infty := \limsup_{t\to\infty} A_t$. The following are equivalent*:

(a) $\mathbb{P}(\bigcup_{t=1}^{\infty} A_t) \le \alpha$.
(b) $\mathbb{P}(A_T) \le \alpha$ *for all random* (*not necessarily stopping*) *times* T.
(c) $\mathbb{P}(A_\tau) \le \alpha$ *for all stopping times* τ, *possibly infinite.*

Our definition of confidence sequences (1.1), based on Darling and Robbins [12] and Lai [43], differs from that Johari, Pekelis and Walsh [35], who require that $\mathbb{P}(\theta_\tau \in \mathrm{CI}_\tau) \ge 1-\alpha$ for all stopping times τ. They allow $\tau = \infty$ by defining $\mathrm{CI}_\infty := \liminf_{t\to\infty} \mathrm{CI}_t$. By taking $A_t := \{\theta_t \notin \mathrm{CI}_t\}$ in Lemma 3, we see that the distinction is immaterial, and furthermore, that we could equivalently define confidence sequences in terms of arbitrary random times, not necessarily stopping times. This generalizes Proposition 1 of Zhao et al. [77].

Always-valid p-values and tests of power one. As an alternative to confidence sequences, Johari, Pekelis and Walsh [35] define an *always-valid p-value process* for some null hypothesis H_0 as an adapted, $[0,1]$-valued sequence $(p_t)_{t=1}^{\infty}$ satisfying $\mathbb{P}_0(p_\tau \le \alpha) \le \alpha$ for all stopping times τ, where $\mathbb{P}_0$ denotes probability under the null H_0. Taking $A_t := \{p_t \le \alpha\}$ in Lemma 3 shows that we may replace this definition with an equivalent one over all random times, not necessarily stopping times, or with the uniform condition $\mathbb{P}_0(\exists t \in \mathbb{N} : p_t \le \alpha) \le \alpha$. By analogy to the usual dual construction between fixed-sample p-values and confidence intervals, one can see that confidence sequences are dual to always-valid p-values, and both are dual to sequential tests, as defined by a stopping time and a binary random variable indicating rejection [35], Proposition 5. In particular, for the null $H_0 : \theta = \theta^\star$, if (CI_t) is a $(1-\alpha)$-confidence sequence for θ, it is clear that a test which stops and rejects the null as soon as $\theta^\star \notin \mathrm{CI}_t$ controls type I error: $\mathbb{P}_0(\text{reject } H_0) = \mathbb{P}_0(\exists t \in \mathbb{N} : \theta^\star \notin \mathrm{CI}_t) \le \alpha$. Typically, then a confidence sequence based on any of the curved uniform bounds in this paper, with radius $u(v) = o(v)$, will yield a *test of power one* [13, 55]. In particular, for a confidence sequence with limits $\bar{X}_t \pm u(V_t)$, it is sufficient that $\bar{X}_t \overset{\text{a.s.}}{\to} \theta$ and $\limsup_{t\to\infty} V_t/t < \infty$ a.s., conditions that usually hold. These conditions imply that the radius of the confidence sequence, $u(V_t)/t$, approaches zero, while the center $\bar{X}_t$ is eventually bounded away from $\theta^\star$ whenever $\theta \ne \theta^\star$, so that the confidence sequence eventually excludes $\theta^\star$ with probability one.

In the one-parameter exponential family case considered in Section 4.4, as noted above, the exponential process $\exp\{\lambda S_t(\mu) - t\psi_\mu(t)\}$ is exactly the likelihood ratio for testing $H_0 : \theta = \theta(\mu)$ against $H_1 : \theta = \theta(\mu) + \lambda$. From the definitions (4.12) and (2), we see that, when using a mixture uniform boundary, a sequential test which rejects as soon as the confidence sequence of Corollary 6 excludes $\mu^\star$ can be seen as equivalently rejecting as soon as either of the mixture likelihood ratios $\int \exp\{\lambda S_t - \psi_{\mu^\star}(\lambda)t\}\,\mathrm{d}F(\lambda)$ or $\int \exp\{-\lambda S_t - \psi_{\mu^\star}(-\lambda)t\}\,\mathrm{d}F(\lambda)$ exceeds $2/\alpha$. Thus a sequential hypothesis test built upon a mixture-based confidence sequence is equivalent to a mixture sequential probability ratio test [55] in the parametric setting. As discussed in Appendix A.6, stitching can be viewed as an approximation to certain mixture bounds, so that hypothesis tests based on stitched bounds are also approximations to mixture SPRTs. Importantly, our confidence sequences are natural nonparametric generalizations of the mixture SPRT, recovering various mixture SPRTs in the parametric settings.

Pros and cons of the running intersection. Our definition (1.1) of a confidence sequence allows for the parameter θ_t to vary with t. It is common in the literature on sequential testing to assume a single, stationary parameter, $\theta_t \equiv \theta$, but this assumption has a troublesome consequence in the context of confidence sequences. If the confidence sequence (CI_t) satisfies $\mathbb{P}(\forall t : \theta \in \mathrm{CI}_t) \geq 1 - \alpha$, then the running intersection $\widetilde{\mathrm{CI}}_t := \bigcap_{s \leq t} \mathrm{CI}_t$ is also uniformly valid for θ, is never larger and may be much smaller. This was observed by Darling and Robbins [13], and is used in the implementation of Johari et al. [34], for example. (In the language of sequential testing, if $(p_t)_{t=1}^{\infty}$ is an always-valid p-value process, then so is $(\min_{s \leq t} p_s)_{t=1}^{\infty}$.)

However, the intersected intervals $\widetilde{\mathrm{CI}}_t$ may become empty at some point. This is particularly likely if the underlying parameter is drifting over time, contrary to the assumption of stationarity or identically distributed observations, and such a drift would be the likely interpretation of this event in practice. In this nonstationary case, the nonintersected sequence is the more sensible one to use. The solution of Johari et al. [34] is to "reset" the experiment, discarding data accumulated up to that point, on the rationale that such an event indicates that previous data are no longer relevant to estimation of the current parameter of interest. However, this means that our confidence sequence can go from a very high precision estimate at some time t to knowing almost nothing at time $t + 1$, which is difficult for an experimenter to interpret and could lead to misleading inference just before the reset. Jennison and Turnbull [32] make a case for the nonintersected intervals on slightly different grounds, arguing that estimation at time t ought to be a function of the sufficient statistic at that time. Shifting to the potential outcomes model in Section 4.2 neatly avoids this issue: because the estimand is changing at each time, the nonintersected intervals are the only reasonable choice for estimating ATE_t and no conceptual difficulty remains.

7. Summary and future work.

We have discussed four techniques for deriving curved uniform boundaries, each improving upon past work, with careful attention paid to constants and to practical issues. By building upon the general framework of Howard et al. [25], we have emphasized the nonparametric applicability of our boundaries. A leading example of the utility of this approach is the general empirical Bernstein bound, with an application to sequential causal inference, and we have also shown how our framework immediately yields novel results for matrix martingales.

7.1. *Other related work.*

We introduced the method of mixtures and the epoch-based analyses in Section 1.1. Two other methods of extending the SPRT deserve mention, though they are distinct from our approaches. First, the approach of Robbins and Siegmund [59, 60] examines $\prod_i f_{\hat{\lambda}_{i-1}}(X_i)/f_0(X_i)$ where $\hat{\lambda}_{i-1}$ is a "nonanticipating" estimate based on $X_1, \ldots, X_{i-1}$. This is similar to a generalized likelihood ratio but modified to retain the martingale property (cf. Wald [74], Section 10.5, [48]). Second, the sequential generalized likelihood ratio approach examines $\sup_\lambda \prod_i f_\lambda(X_i)/f_0(X_i)$, which is not a martingale under the null [40, 44, 66].

The concept of *test (super)martingales* expounded by Shafer et al. [63] is related to our methods for conducting inference based on Ville's inequality applied to nonnegative supermartingales. Their main example is the Beta mixture for i.i.d. Bernoulli observations, an example which originated with Ville [72] and discussed by Robbins [55] and Lai [41]. A recent "safe testing" framework of Grünwald, de Heide and Koolen [23] is also tightly related. In terms of these frameworks, our work can be viewed as constructing "safe confidence intervals" (and thus safe tests) using nonparametric test supermartingales.

A very different approach is that of group sequential methods [33, 47, 52, 53]. These methods rely on either exact discrete distributions or asymptotics to assume exact normality of group increments, either of which permits computation of sequential boundaries via numerical integration. The resulting confidence sequences are tighter than ours, but lack nonasymptotic guarantees or closed-form results and do not support continuous monitoring.

A related problem is that of terminal confidence intervals, in which one assumes a rigid stopping rule and wishes to construct a confidence interval upon termination. Siegmund [64] gave an analytical treatment of the problem; numerical methods are also available for group sequential tests [33], Section 8.5. However, the idea of a rigid stopping rule is often restrictive.

7.2. *Future work.*

We discuss in Appendix I how our work may be extended to martingales in smooth Banach spaces and real-valued, continuous-time martingales. It may be fruitful to explore applications in those areas.

Our consideration of optimality has been limited to the discussion in Section 3.6. It would be valuable to further explore various optimality properties for nonasymptotic uniform bounds. For example, it is standard in sequential testing to compute the expected sample size to reject a null under parametric alternatives. Though we target less restrictive assumptions, it may be instructive to compute bounds in special cases. Second, a natural counterpoint to our uniform concentration bounds would be a set of uniform anticoncentration bounds. This would yield a nonasymptotic extension of the "lim inf" half of the classical LIL. Balsubramani [5], Theorem 3, gives one such interesting result. Last, in practice, one will rarely require updated inference after every observation, and may be content to take observations in groups. Further, one may be satisfied with a finite time horizon [21]. This is the domain in which group-sequential methods shine, but SPRT-based methods can be made competitive by estimating the "overshoot" of the stopped supermartingale [45, 46, 65, 75]. It would be interesting to understand whether such improvements work out in nonparametric settings.

Acknowledgments. We thank Boyan Duan for catching typos, and the referees and Associate Editor for useful suggestions.

The first author was supported by Office of Naval Research grant N00014-17-1-2176.

The second author was supported by NSF grant DMS1916320.

The fourth author was supported by Office of Naval Research grants N00014-15-1-2367, N00014-17-1-2176.

Jon McAuliffe is also with The Voleon Group.

SUPPLEMENTARY MATERIAL

Supplement to "Time-uniform, nonparametric, nonasymptotic confidence sequences" (DOI: 10.1214/20-AOS1991SUPP; .pdf). Proofs, additional figures, implementation details, and extension to smooth Banach spaces and continuous-time processes.

REFERENCES

[1] ARMITAGE, P., MCPHERSON, C. K. and ROWE, B. C. (1969). Repeated significance tests on accumulating data. *J. Roy. Statist. Soc. Ser. A* **132** 235–244. MR0250405 https://doi.org/10.2307/2343787

[2] ARONOW, P. M. and MIDDLETON, J. A. (2013). A class of unbiased estimators of the average treatment effect in randomized experiments. *J. Causal Inference* **1** 135–154.

[3] AUDIBERT, J.-Y., MUNOS, R. and SZEPESVÁRI, C. (2009). Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoret. Comput. Sci.* **410** 1876–1902. MR2514714 https://doi.org/10.1016/j.tcs.2009.01.016

[4] AZUMA, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Math. J.* (2) **19** 357–367. MR0221571 https://doi.org/10.2748/tmj/1178243286

[5] BALSUBRAMANI, A. (2014). Sharp finite-time iterated-logarithm martingale concentration. arXiv:1405.2639.

[6] BALSUBRAMANI, A. and RAMDAS, A. (2016). Sequential nonparametric testing with the law of the iterated logarithm. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*. *UAI'*16 42–51. AUAI Press.

[7] BERCU, B., DELYON, B. and RIO, E. (2015). *Concentration Inequalities for Sums and Martingales*. *SpringerBriefs in Mathematics*. Springer, Cham. MR3363542 https://doi.org/10.1007/978-3-319-22099-4

[8] BERMAN, R., PEKELIS, L., SCOTT, A. and VAN DEN BULTE, C. (2018). p-hacking and false discovery in A/B testing. Technical Report No. 3204791. SSRN.

[9] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities*: *A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. With a foreword by Michel Ledoux. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

[10] CHERNOFF, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann*. *Math*. *Stat*. **23** 493–507. MR0057518 https://doi.org/10.1214/aoms/1177729330

[11] CRAMÉR, H. (1938). Sur un nouveau théorème-limite de la théorie des probabilités. *Actualités Scientifiques* **736**.

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

[13] DARLING, D. A. and ROBBINS, H. (1967). Iterated logarithm inequalities. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **57** 1188–1192. MR0211441 https://doi.org/10.1073/pnas.57.5.1188

[14] DARLING, D. A. and ROBBINS, H. (1968). Some further remarks on inequalities for sample sums. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **60** 1175–1182. MR0235604 https://doi.org/10.1073/pnas.60.4.1175

[15] DE LA PEÑA, V. H., KLASS, M. J. and LAI, T. L. (2004). Self-normalized processes: Exponential inequalities, moment bounds and iterated logarithm laws. *Ann*. *Probab*. **32** 1902–1933. MR2073181 https://doi.org/10.1214/009117904000000397

[16] DE LA PEÑA, V. H., KLASS, M. J. and LAI, T. L. (2007). Pseudo-maximization and self-normalized processes. *Probab*. *Surv*. **4** 172–192. MR2368950 https://doi.org/10.1214/07-PS119

[17] DE LA PEÑA, V. H., KLASS, M. J. and LAI, T. L. (2009). Theory and applications of multivariate self-normalized processes. *Stochastic Process*. *Appl*. **119** 4210–4227. MR2565565 https://doi.org/10.1016/j.spa.2009.10.003

[18] DE LA PEÑA, V. H., LAI, T. L. and SHAO, Q.-M. (2009). *Self-Normalized Processes*: *Limit Theory and Statistical Applications*. *Probability and Its Applications* (*New York*). Springer, Berlin. MR2488094 https://doi.org/10.1007/978-3-540-85636-8

[19] EFRON, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika* **58** 403–417. MR0312660 https://doi.org/10.1093/biomet/58.3.403

[20] GARIVIER, A. (2013). Informational confidence bounds for self-normalized averages and applications. In 2013 *IEEE Information Theory Workshop* (*ITW*) 1–5. IEEE.

[21] GARIVIER, A. and LEONARDI, F. (2011). Context tree selection: A unifying view. *Stochastic Process*. *Appl*. **121** 2488–2506. MR2832411 https://doi.org/10.1016/j.spa.2011.06.012

[22] GITTENS, A. and TROPP, J. A. (2011). Tail bounds for all eigenvalues of a sum of random matrices. ACM Report 2014-02, Caltech.

[23] GRÜNWALD, P., DE HEIDE, R. and KOOLEN, W. (2019). Safe testing. arXiv:1906.07801.

[24] HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J*. *Amer*. *Statist*. *Assoc*. **58** 13–30. MR0144363

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

[26] HOWARD, S. R., RAMDAS, A., MCAULIFFE, J. and SEKHON, J. (2021). Supplement to "Time-uniform, nonparametric, nonasymptotic confidence sequences." https://doi.org/10.1214/20-AOS1991SUPP

[27] IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—for Statistics*, *Social*, *and Biomedical Sciences*: *An Introduction*. Cambridge Univ. Press, New York. MR3309951 https://doi.org/10.1017/CBO9781139025751

[28] JAMIESON, K. and JAIN, L. (2018). A bandit approach to multiple testing with false discovery control. In *Proceedings of the* 32*nd International Conference on Neural Information Processing Systems* 3664–3674.

[29] JAMIESON, K., MALLOY, M., NOWAK, R. and BUBECK, S. (2014). lil' UCB: An optimal exploration algorithm for multi-armed bandits. In *Proceedings of the* 27*th Conference on Learning Theory* **35** 423–439.

[30] JAMIESON, K. and NOWAK, R. (2014). Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In 48*th Annual Conference on Information Sciences and Systems* (*CISS*) 1–6.

[31] JENNISON, C. and TURNBULL, B. W. (1984). Repeated confidence intervals for group sequential clinical trials. *Control*. *Clin*. *Trials* **5** 33–45.

[32] JENNISON, C. and TURNBULL, B. W. (1989). Interim analyses: The repeated confidence interval approach. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *B* **51** 305–361. With discussion and a reply by the authors. MR1017201

[33] JENNISON, C. and TURNBULL, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. CRC Press/CRC, Boca Raton, FL. MR1710781

[34] JOHARI, R., KOOMEN, P., PEKELIS, L. and WALSH, D. (2017). Peeking at A/B tests: Why it matters, and what to do about it. 1517–1525. ACM Press.

[35] JOHARI, R., PEKELIS, L. and WALSH, D. J. (2015). Always valid inference: Bringing sequential analysis to A/B testing. arXiv preprint arXiv:1512.04922.

[36] JØRGENSEN, B. (1997). *The Theory of Dispersion Models*. *Monographs on Statistics and Applied Probability* **76**. CRC Press, London. MR1462891

[37] KAUFMANN, E., CAPPÉ, O. and GARIVIER, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *J*. *Mach*. *Learn*. *Res*. **17** Paper No. 1, 42. MR3482921

[38] KAUFMANN, E. and KOOLEN, W. (2018). Mixture martingales revisited with applications to sequential tests and confidence intervals. arXiv:1811.11419.

[39] KOLTCHINSKII, V. and LOUNICI, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133. MR3556768 https://doi.org/10.3150/15-BEJ730

[40] KULLDORFF, M., DAVIS, R. L., KOLCZAK, M., LEWIS, E., LIEU, T. and PLATT, R. (2011). A maximized sequential probability ratio test for drug and vaccine safety surveillance. *Sequential Anal*. **30** 58–78. MR2770706 https://doi.org/10.1080/07474946.2011.539924

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

[42] LAI, T. L. (1976). Boundary crossing probabilities for sample sums and confidence sequences. *Ann*. *Probab*. **4** 299–312. MR0405578 https://doi.org/10.1214/aop/1176996135

[43] LAI, T. L. (1984). Incorporating scientific, ethical and economic considerations into the design of clinical trials in the pharmaceutical industry: A sequential approach. *Comm*. *Statist*. *Theory Methods* **13** 2355–2368.

[44] LAI, T. L. (1997). On optimal stopping problems in sequential hypothesis testing. *Statist*. *Sinica* **7** 33–51. MR1441143

[45] LAI, T. L. and SIEGMUND, D. (1977). A nonlinear renewal theory with applications to sequential analysis. I. *Ann*. *Statist*. **5** 946–954. MR0445599

[46] LAI, T. L. and SIEGMUND, D. (1979). A nonlinear renewal theory with applications to sequential analysis. II. *Ann*. *Statist*. **7** 60–76. MR0515684

[47] LAN, K. K. G. and DEMETS, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70** 659–663. MR0725380 https://doi.org/10.2307/2336502

[48] LORDEN, G. and POLLAK, M. (2005). Nonanticipating estimation applied to sequential analysis and changepoint detection. *Ann*. *Statist*. **33** 1422–1454. MR2195641 https://doi.org/10.1214/009053605000000183

[49] MALEK, A., KATARIYA, S., CHOW, Y. and GHAVAMZADEH, M. (2017). Sequential multiple hypothesis testing with type I error control. In *Artificial Intelligence and Statistics* 1468–1476.

[50] MAURER, A. and PONTIL, M. (2009). Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the Conference on Learning Theory*.

[51] MCDIARMID, C. (1998). Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics*. *Algorithms Combin*. **16** 195–248. Springer, Berlin. MR1678578 https://doi.org/10.1007/978-3-662-12788-9_6

[52] O'BRIEN, P. C. and FLEMING, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35** 549–556.

[53] POCOCK, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64** 191–199.

[54] RAGINSKY, M., SASON, I. et al. (2013). Concentration of measure inequalities in information theory, communications, and coding. *Found*. *Trends Commun*. *Inf*. *Theory* **10** 1–246.

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

[56] ROBBINS, H. and SIEGMUND, D. (1968). Iterated logarithm inequalities and related statistical procedures. In *Mathematics of the Decision Sciences*, *Part* 2 (*Seminar*, *Stanford Calif*., 1967) 267–279. Amer. Math. Soc., Providence, RI. MR0251777

[57] ROBBINS, H. and SIEGMUND, D. (1969). Probability distributions related to the law of the iterated logarithm. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **62** 11–13. MR0242228 https://doi.org/10.1073/pnas.62.1.11

[58] ROBBINS, H. and SIEGMUND, D. (1970). Boundary crossing probabilities for the Wiener process and sample sums. *Ann*. *Math*. *Stat*. **41** 1410–1429. MR0277059 https://doi.org/10.1214/aoms/1177696787

[59] ROBBINS, H. and SIEGMUND, D. (1972). A class of stopping rules for testing parametric hypotheses. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California*, *Berkeley*, *Calif.*, 1970/1971), *Vol. IV*: *Biology and Health* 37–41. MR0403111

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

[61] RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.

[62] RUDELSON, M. (1999). Random vectors in the isotropic position. *J. Funct. Anal.* **164** 60–72. MR1694526 https://doi.org/10.1006/jfan.1998.3384

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

[64] SIEGMUND, D. (1978). Estimation following sequential tests. *Biometrika* **65** 341–349. MR0513934 https://doi.org/10.2307/2335213

[65] SIEGMUND, D. (1985). *Sequential Analysis*: *Tests and Confidence Intervals*. *Springer Series in Statistics*. Springer, New York. MR0799155 https://doi.org/10.1007/978-1-4757-1862-1

[66] SIEGMUND, D. and GREGORY, P. (1980). A sequential clinical trial for testing $p_1 = p_2$. *Ann. Statist.* **8** 1219–1228. MR0594639

[67] SPLAWA-NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. Translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. MR1092986

[68] STOUT, W. F. (1970). The Hartman–Wintner law of the iterated logarithm for martingales. *Ann. Math. Stat.* **41** 2158–2160.

[69] TROPP, J. A. (2011). Freedman's inequality for matrix martingales. *Electron. Commun. Probab.* **16** 262–270. MR2802042 https://doi.org/10.1214/ECP.v16-1624

[70] TROPP, J. A. (2015). An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.* **8** 1–230.

[71] VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

[72] VILLE, J. (1939). *Étude Critique de la Notion de Collectif*. NUMDAM. MR3533075

[73] WALD, A. (1945). Sequential tests of statistical hypotheses. *Ann. Math. Stat.* **16** 117–186. MR0013275 https://doi.org/10.1214/aoms/1177731118

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

[75] WHITEHEAD, J. and STRATTON, I. (1983). Group sequential clinical trials with triangular continuation regions. *Biometrics* **39** 227–236. MR0712749 https://doi.org/10.2307/2530822

[76] YANG, F., RAMDAS, A., JAMIESON, K. G. and WAINWRIGHT, M. J. (2017). A framework for Multi-A(rmed)/B(andit) testing with online FDR control. In 31*st Conference on Neural Information Processing Systems*.

[77] ZHAO, S., ZHOU, E., SABHARWAL, A. and ERMON, S. (2016). Adaptive concentration inequalities for sequential decision problems. In 30*th Conference on Neural Information Processing Systems*.