Time-uniform, nonparametric, nonasymptotic confidence sequences

Steven R. Howard[1] Aaditya Ramdas[2,3] Jon McAuliffe[1,4] Jasjeet Sekhon[1,5]
Departments of Statistics[1] and Political Science[5], UC Berkeley
Departments of Statistics and Data Science[2] and Machine Learning[3], Carnegie Mellon
The Voleon Group[4]
{stevehoward,jonmcauliffe,sekhon}@berkeley.edu, aramdas@stat.cmu.edu

August 9, 2022

Abstract

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 (Armitage et al., 1969; Berman et al., 2018). 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 (Darling and Robbins, 1967*a*; Lai, 1984; Jennison and Turnbull, 1989). 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 t^{th} unit, we calculate an updated confidence set CI_t for the unknown quantity of interest θ_t, with the uniform coverage property

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

With only a uniform lower bound (L_t), i.e., 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. (2020), 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 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.

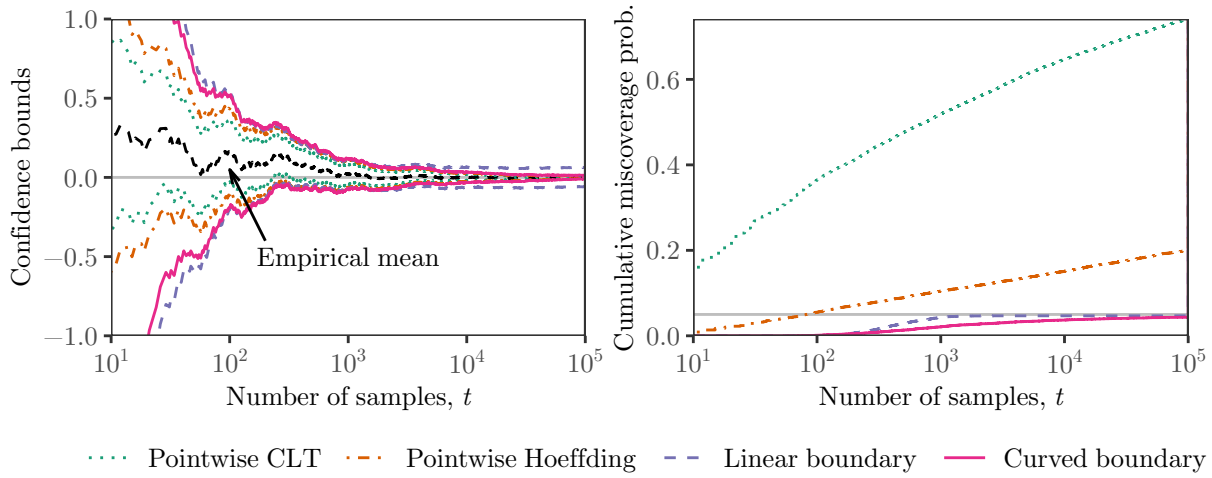

Figure 1: Left panel shows 95% pointwise confidence intervals and uniform confidence sequences for the mean of a Rademacher random variable, using one simulation of 100,000 i.i.d. draws. Right panel shows cumulative chance of miscoverage based on 10,000 replications; flat grey line shows the nominal 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, (14), qualitatively similar to the stitched bound (2).

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 (10):

$$\frac{\sum_{i=1}^{t} X_i}{t} \pm 1.7\sqrt{\frac{\log\log(2t) + 0.72\log(10.4/\alpha)}{t}}. \tag{2}$$

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* (Jamieson et al., 2014).

We develop confidence sequences that possess the following properties:

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

(P2) **Unbounded sample size**: our methods 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 towards 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 9 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 (1967*a*). They are called *repeated confidence intervals* by Jennison and Turnbull (1984, 1989) (with a focus on finite time horizons) and *always-valid confidence intervals* by Johari et al. (2015). They are sometimes labeled *anytime confidence intervals* in the machine learning literature (Jamieson and Jain, 2018).

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 (Johari et al., 2015). 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) 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) 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 (2013), Jamieson et al. (2014), Kaufmann et al. (2016), and Zhao et al. (2016) 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 (Jamieson and Nowak, 2014) and related methods for sequential testing with multiple comparisons (Yang et al., 2017; Malek et al., 2017; Jamieson and Jain, 2018). Our results improve and generalize such methods.

Maurer and Pontil (2009) and Audibert et al. (2009) prove empirical-Bernstein bounds for fixed times or finite time horizons. Our empirical-Bernstein bound holds uniformly over infinite time. Balsubramani (2014) 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 (2016) 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 pre-factor 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, (Wald, 1945)), such as in Section 3.6 of Howard et al. (2020). Wald worked in a parametric setting, though it is known that the normal SPRT depends only on sub-Gaussianity (e.g., Robbins (1970)). The resulting confidence sequence does not shrink towards 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 (Ville, 1939; Wald, 1945; Darling and Robbins, 1968; Robbins and Siegmund, 1969, 1970; Robbins, 1970; Lai, 1976*b*; de la Peña et al., 2007; Balsubramani, 2014; Bercu et al., 2015; Kaufmann and Koolen, 2018). 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 (Darling and Robbins, 1967*b*; Robbins and Siegmund, 1968; Kaufmann et al., 2016).

The literature on self-normalized bounds makes extensive use of the method of mixtures, sometimes called pseudo-maximization (de la Peña et al., 2004, 2007; de la Peña, Klass and Lai, 2009; de la Peña, Lai and Shao, 2009; Garivier, 2013); 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 et al. (2004, Eq. 4.20) includes an infinite-horizon curve-crossing bound. Lai (1976*b*) 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. (2017) 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 sub-exponential 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 sub-exponential 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 $\mu_t \coloneqq t^{-1} \sum_{i=1}^{t} \mathbb{E}_{i-1} X_i$ at each time t using the sample mean $\bar{X}_t \coloneqq t^{-1} \sum_{i=1}^{t} X_i$; 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 $S_t \coloneqq \sum_{i=1}^{t} (X_i - \mathbb{E}_{i-1} X_i)$, 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

$$\mathbb{P}\big(\exists t \geq 1 : S_t \geq u_\alpha(V_t)\big) \leq \alpha \tag{3}$$

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

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).

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 (3), we build upon the following general condition (Howard et al., 2020, 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

$$\exp\{\lambda S_t - \psi(\lambda) V_t\} \leq L_t(\lambda) \text{ a.s. for all } t. \tag{4}$$

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):

$$\mathbb{S}_\psi^{l_0} := \{(S_t, V_t) : (S_t) \text{ is } l_0\text{-sub-}\psi \text{ with variance process } (V_t)\}. \tag{5}$$

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 ((Hoeffding, 1963; Azuma, 1967; McDiarmid, 1998); (Raginsky et al., 2013, section 2.2)), itself based on the classical Cramér-Chernoff method ((Cramér, 1938; Chernoff, 1952); (Boucheron et al., 2013, 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 $\lambda \geq 0$ to derive 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, i.e., 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 3 and 4, drawn from Howard et al. (2020), 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 $\log \mathbb{E} e^{\lambda X_t} \leq \psi(\lambda)$ for all $t \geq 1$ and $\lambda \in \mathbb{R}$, then writing $S_t = \sum_{i=1}^t X_i$, we have $\mathbb{P}(S_t \geq x) \leq e^{-t\psi^\star(x/t)}$ 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 $\mathbb{P}(S_t \geq z_\alpha(t)) \leq \alpha$ 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 $l_0 \geq 1$, a function $u : \mathbb{R} \to \mathbb{R}$ is called an *l_0-sub-ψ uniform boundary* with crossing probability α if

$$\sup_{(S_t, V_t) \in \mathbb{S}_\psi^{l_0}} \mathbb{P}(\exists t \geq 1 : S_t \geq u(V_t)) \leq \alpha. \tag{6}$$

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:

- $\psi_{B,g,h}(\lambda) \coloneqq \frac{1}{gh}\log\left(\frac{ge^{h\lambda}+he^{-g\lambda}}{g+h}\right)$ 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 $g + h = 1$.
- $\psi_N(\lambda) \coloneqq \lambda^2/2$ on $0 \le \lambda < \infty$, the CGF of a standard Gaussian r.v.
- $\psi_{P,c}(\lambda) \coloneqq c^{-2}(e^{c\lambda} - c\lambda - 1)$ 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 $\psi_{P,0} = \psi_N$.
- $\psi_{E,c}(\lambda) \coloneqq c^{-2}(-\log(1 - c\lambda) - c\lambda)$ 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 $\psi_{E,0} = \psi_N$.
- $\psi_{G,c}(\lambda) \coloneqq \lambda^2/(2(1 - c\lambda))$ 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 et al. (2013). 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 below.

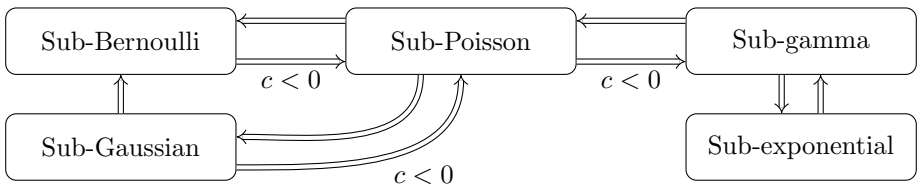

Figure 2: Relations among sub-ψ boundaries: each arrow indicates that a sub-ψ boundary at the source node can also serve as a sub-ψ boundary at the destination node, with appropriate modifications to their parameters. Details are in Proposition 11.

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 11, 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 sub-exponential boundary can serve as a sub-ψ boundary in any of the other four cases. Indeed, a sub-gamma or sub-exponential boundary applies to many cases of practical interest, as detailed below.

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 sub-exponential 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. (2020, Proposition 1), which shows that any process (S_t) which is sub-ψ is also sub-gamma and sub-exponential, 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 (Jorgensen, 1997, 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 $S_t \coloneqq \sigma^{-2}\sum_{i=1}^{t+1}(X_i - \bar{X}_{t+1})^2 - t$ for $t = 1, 2, \ldots$, where $\bar{X}_t \coloneqq t^{-1}\sum_{i=1}^{t} X_i$ is the sample mean. This S_t is a centered and scaled sample variance, and as in Darling and Robbins (1967*a*), 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-sub-exponential with variance process $V_t = 2t$ and scale parameter $c = 2$. We may uniformly bound the upper deviations of S_t using any sub-exponential uniform boundary, for example the gamma-exponential mixture boundary of Proposition 9. Or, we can use Proposition 11 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-sub-exponential with scale parameter $c = -2$. Now Proposition 11 implies that $(-S_t)$ is sub-gamma with scale $c = -1$, so the stitched boundary again applies, while Proposition 11 implies that $(-S_t)$ is also sub-Gaussian, so we may alternatively use the normal mixture boundary of Proposition 6. 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 (1967*a*).

The simplest uniform boundaries are linear with positive intercept and slope. This is formalized in Howard et al. (2020), partially restated below.

Lemma 1 ((Howard et al., 2020), Theorem 1). *For any $\lambda \in [0, \lambda_{\max})$ and $\alpha \in (0, 1)$,*

$$u(v) := \frac{\log(l_0/\alpha)}{\lambda} + \frac{\psi(\lambda)}{\lambda} \cdot v \tag{7}$$

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

$$\mathcal{S}_\alpha(v) := \sqrt{k_1^2 v\ell(v) + k_2^2 c^2 \ell^2(v)} + k_2 c\ell(v), \text{ where } \begin{cases} \ell(v) := \log h(\log_\eta(\frac{v}{m})) + \log(\frac{l_0}{\alpha}), \\ k_1 := (\eta^{1/4} + \eta^{-1/4})/\sqrt{2}, \\ k_2 := (\sqrt{\eta} + 1)/2, \end{cases} \tag{8}$$

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 the examples we consider, $\ell(v)$ 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 $V_t/\ell(V_t) \gg 2c^2\sqrt{\eta}$.

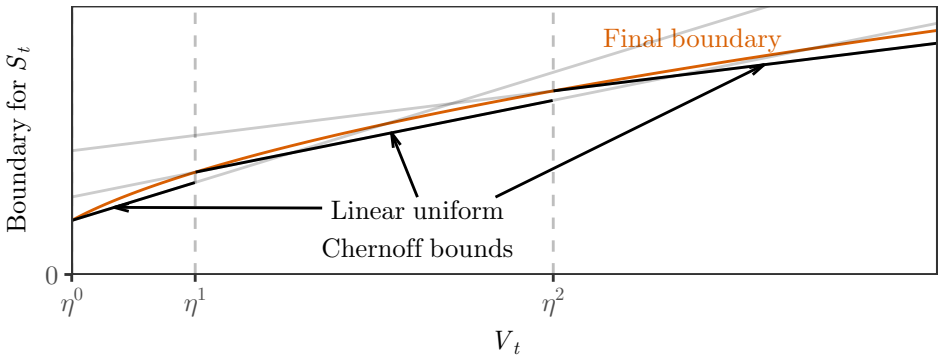

Figure 3: Illustration of Theorem 1, stitching together linear boundaries to construct a curved boundary. We break time into geometrically-spaced epochs $\eta^k \le V_t < \eta^{k+1}$, construct a linear uniform bound using Lemma 1 optimized for each epoch, and take a union bound over all crossing events.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

Theorem 1 (Stitched boundary)**.** *For any $c \ge 0, \alpha \in (0,1), \eta > 1, m > 0$, and $h : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ increasing such that $\sum_{k=0}^{\infty} 1/h(k) \le 1$, 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 $v_0 \ge m$,*

$$\mathbb{P}\left(\exists t \ge 1 : V_t \ge v_0 \text{ and } S_t \ge \mathcal{S}_\alpha(V_t)\right) \le \sum_{k=\lfloor \log_\eta(v_0/m) \rfloor}^{\infty} \frac{\alpha}{h(k)}. \tag{9}$$

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 (9) implies $\mathbb{P}\left(\sup_t V_t = \infty \text{ and } S_t \ge \mathcal{S}_\alpha(V_t) \text{ infinitely often}\right) = 0$. 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 k^{th} epoch to equal $\alpha/h(k)$. Then Theorem 1 gives a curved boundary which grows at a rate $\mathcal{O}\left(\sqrt{V_t \log h(\log_\eta V_t)}\right)$ 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 $c \ge 0$,

$$\mathcal{S}_\alpha(v) = k_1 \sqrt{v\left(s \log\log\left(\frac{\eta v}{m}\right) + \log \frac{\zeta(s)}{\alpha \log^s \eta}\right)} + c k_2 \left(s \log\log\left(\frac{\eta v}{m}\right) + \log \frac{\zeta(s)}{\alpha \log^s \eta}\right), \tag{10}$$

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 (Stout, 1970). 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

$$\mathbb{P}\left(\exists t \geq 1 : S_t \geq 1.7\sqrt{t\left(\log\log(2t) + 0.72\log\left(\frac{5.2}{\alpha}\right)\right)}\right) \leq \alpha. \tag{11}$$

Figure 9 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 10 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 sub-exponential 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 8.

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*

$$\limsup_{t\to\infty} \frac{S_t}{\sqrt{2V_t \log\log V_t}} \leq 1 \quad \text{on} \quad \left\{\sup_t V_t = \infty\right\}. \tag{12}$$

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 (Ville, 1939; Wald, 1945; Darling and Robbins, 1968; Robbins, 1970; Robbins and Siegmund, 1969, 1970; Lai, 1976*b*; Kaufmann and Koolen, 2018). 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 $\alpha \in (0,1)$,*

$$\mathcal{M}_\alpha(v) := \sup\left\{s \in \mathbb{R} : \underbrace{\int \exp\{\lambda s - \psi(\lambda)v\}\,\mathrm{d}F(\lambda)}_{=:m(s,v)} < \frac{l_0}{\alpha}\right\} \tag{13}$$

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 (sub-exponential case).

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

$$u(v) := \sqrt{(v+\rho)\log\left(\frac{l_0^2(v+\rho)}{\alpha^2\rho}\right)}. \tag{14}$$

The one-sided normal mixture boundary has a similar, closed-form upper bound, making these especially convenient. It is clear from (14) 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 (13) 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*

$$\mathcal{M}_\alpha(v) = \sqrt{v\left[c\log\left(\frac{cl_0^2 v}{2\pi\alpha^2 f^2(0)}\right) + o(1)\right]} \quad \text{as } v\to\infty. \tag{15}$$

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 (1970, Section 4) and Lai (1976*a*, 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 9). 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 $(0,\overline{\lambda}]$ for some $\overline{\lambda}\in(0,\lambda_{\max})$. 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:

$$\lambda_k := \frac{\overline{\lambda}}{\eta^{k+1/2}} \quad \text{and} \quad w_k := \frac{\overline{\lambda}(\eta-1)f(\lambda_k\sqrt{\eta})}{\eta^{k+1}} \quad \text{for} \quad k = 0,1,2,\ldots. \tag{16}$$

Theorem 2 (Discrete mixture bound). *Fix $\psi : [0, \lambda_{\max}) \to \mathbb{R}$, $\alpha \in (0, 1)$, $\overline{\lambda} \in (0, \lambda_{\max})$, and a probability density f on $(0, \overline{\lambda}]$ that is nonincreasing and positive. For supports λ_k and weights w_k defined in (16),*

$$\mathrm{DM}_\alpha(v) \coloneqq \sup\left\{ s \in \mathbb{R} : \sum_{k=0}^{\infty} w_k \exp\left\{\lambda_k s - \psi(\lambda_k) v\right\} < \frac{l_0}{\alpha} \right\}, \tag{17}$$

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

We suppress the dependence of DM_α on f, l_0, $\overline{\lambda}$ and η for notational simplicity. Though Theorem 2 is a straightforward consequence of the method of mixtures, our choice of discretization (16) makes it effective, broadly applicable, and easy to implement. See Appendix A.5 for the proof of this result. Figure 9 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 (16) 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*

$$u(v) \coloneqq \begin{cases} g(1 \vee v), & v \le v_{\max}, \\ \infty, & \text{otherwise} \end{cases} \tag{18}$$

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

$$l_0 \inf_{\eta > 1} \sum_{k=0}^{\lceil \log_\eta v_{\max} \rceil} \exp\left\{ -\frac{2(g(\eta^{k+1}) - g(\eta^k))(\eta g(\eta^k) - g(\eta^{k+1}))}{\eta^k (\eta - 1)^2} \right\}. \tag{19}$$

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 (19) no longer applies. A similar idea was considered by Darling and Robbins (1968), 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

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

This boundary is illustrated in Figure 9.

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 1,000 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.

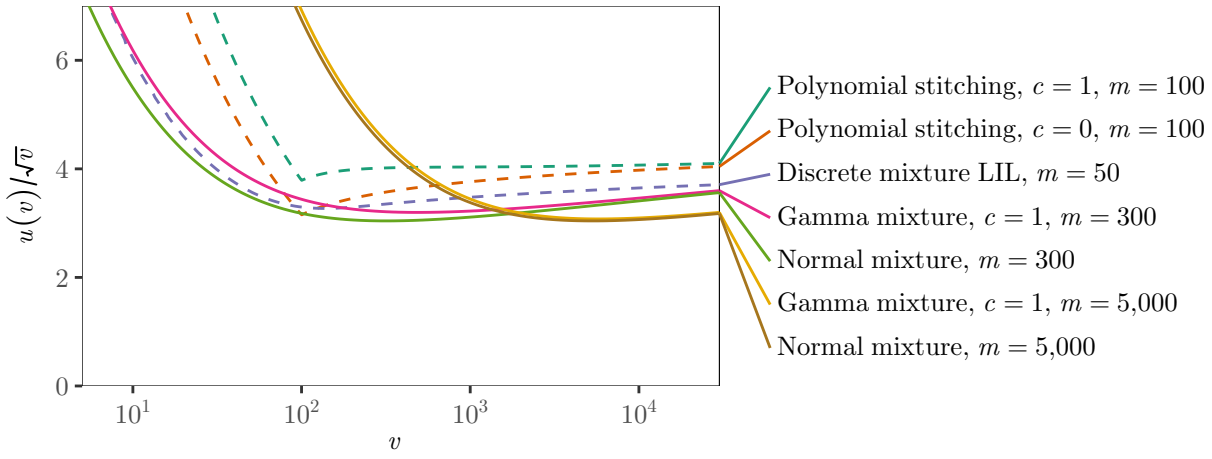

Figure 4: Comparison of normalized uniform boundaries $u(v)/\sqrt{v}$ optimized for different intrinsic times. Normal mixture uses Appendix Proposition 6, while gamma mixture uses Appendix Proposition 9. Polynomial stitched boundary is given in (10), with $\eta = 2$ and $s = 1.4$. Discrete mixture applies 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 $\alpha = 0.025$.

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 (14), 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 (14) 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

$$\frac{m}{\rho} = -W_{-1}\left(-\frac{\alpha^2}{el_0^2}\right) - 1. \tag{21}$$

(b) For fixed $m > 0$, the choice of ρ which minimizes the boundary value $u(m)$ is also determined by (21).

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 $m = 5{,}000$. 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.

The one-sided normal mixture boundary of Appendix Proposition 6 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 upper bound on crossing probability 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 9 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 (1970), 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 $\mathcal{M}_\alpha$,*

$$\sup_{(S_t,V_t)\in\mathbb{S}^1_{\psi_N}} \mathbb{P}(\exists t \geq 1 : S_t \geq \mathcal{M}_\alpha(V_t)) = \alpha. \tag{22}$$

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 uniform boundaries 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 sub-exponential uniform boundary with crossing probability α for scale $c = b-a$. Then*

$$\mathbb{P}\left(\forall t \geq 1 : \left|\bar{X}_t - \mu_t\right| < \frac{u\left(\sum_{i=1}^t (X_i - \widehat{X}_i)^2\right)}{t}\right) \geq 1 - 2\alpha. \tag{23}$$

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 (23) 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 $\widehat{V}_t \coloneqq \sum_{i=1}^t (X_i - \bar{X}_{i-1})^2$. Invoking Theorem 4 with the polynomial stitched bound (10) using $c=1$, $\eta = 2$, $m=1$, and $h(k) \propto k^{1.4}$, we have the following 95%-confidence sequence for μ_t:

$$\bar{X}_t \pm \frac{1.7\sqrt{(\widehat{V}_t \vee 1)(\log\log(2(\widehat{V}_t \vee 1)) + 3.8)} + 3.4\log\log(2(\widehat{V}_t \vee 1)) + 13}{t}. \tag{24}$$

When a closed form is not required, the gamma-exponential mixture (Proposition 9) 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 (Neyman, 1923/1990; Rubin, 1974; Imbens and Rubin, 2015). 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 $Y_t^{\text{obs}} \coloneqq Y_t(Z_t)$. 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 $P_t \coloneqq E_{t-1} Z_t$ 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 Efron (1971) and related covariate balancing methods.

At each step t, having treated and observed units $1,\ldots,t$, we wish to draw inference about the estimand $\text{ATE}_t \coloneqq t^{-1}\sum_{i=1}^t [Y_i(1) - Y_i(0)]$. 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

$$X_t \coloneqq \widehat{Y}_t(1) - \widehat{Y}_t(0) + \left(\frac{Z_t - P_t}{P_t(1-P_t)}\right)(Y_t^{\text{obs}} - \widehat{Y}_t(Z_t)), \tag{25}$$

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 (2013) 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

$$V_t \coloneqq \sum_{i=1}^t (X_i - \widehat{X}_i)^2 = \sum_{i=1}^t \left(\frac{Z_i - P_i}{P_i(1-P_i)}\right)^2 (Y_i^{\text{obs}} - \widehat{Y}_i(Z_i))^2. \tag{26}$$

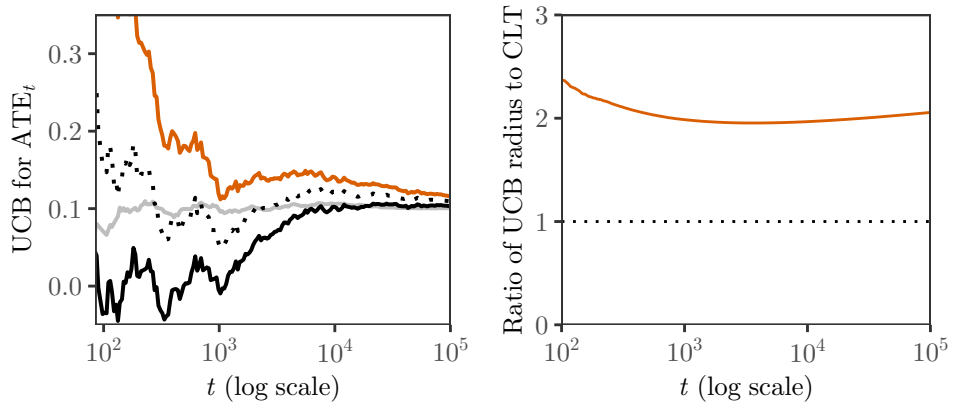

Figure 5: Upper half of 95% empirical-Bernstein confidence sequence for ATE_t under Bernoulli randomization based on one simulated sequence of i.i.d. observations, $P_t \equiv 0.5$, $Y_i(0) \sim \mathrm{Ber}(0.5)$, $Y_i(1) = \xi_i \vee Y_i(0)$ where $\xi_i \sim \mathrm{Ber}(0.2)$. Grey line shows estimand 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 $\widehat{Y}_t(k) = \sum_{i=1}^{t-1} Y_i^{\mathrm{obs}} 1_{Z_i=k} / \sum_{i=1}^{t-1} 1_{Z_i=k}$, $\alpha = 0.05$ and a gamma-exponential mixture bound with $\rho = 12.6$, chosen to optimize for intrinsic time $V_t = 100$.

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 sub-exponential uniform boundary with scale $2/p_{\min}$ and crossing probability α. Then*

$$\mathbb{P}\left(\forall t \geq 1 : \left|\bar{X}_t - \mathrm{ATE}_t\right| < \frac{u(V_t)}{t}\right) \geq 1 - 2\alpha. \tag{27}$$

For u, one may choose the gamma-exponential mixture boundary (Proposition 9) or the stitched boundary (10) 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.

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. (2020, 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 (8).

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 $V_t \coloneqq \gamma_{\max}(\sum_{i=1}^{t} \mathbb{E}_{t-1}(Y_t - Y_{t-1})^2)$ and $S_t \coloneqq \gamma_{\max}(Y_t)$. Then for any $\eta > 1, s > 1, m > 0, \alpha \in (0,1)$, we have*

$$\mathbb{P}\left(\exists t \geq 1 : S_t \geq k_1(\eta)\sqrt{(V_t \vee m)\ell_{\eta,s}(V_t \vee m)} + \frac{b k_2(\eta)}{3}\ell_{\eta,s}(V_t \vee m)\right) \leq \alpha. \tag{28}$$

The result follows using the polynomial stitched boundary after invoking Fact 1(c) and Lemma 2 of Howard et al. (2020) (cf. (Tropp, 2011)), 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 $\Delta Y_t \coloneqq Y_t - Y_{t-1}$.

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

$$\limsup_{t\to\infty} \frac{\gamma_{\max}(Y_t)}{\sqrt{2V_t \log\log V_t}} \le 1 \quad \text{a.s. on } \left\{\sup_t V_t = \infty\right\} \tag{29}$$

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 $k > 2$, $\mathbb{E}_{t-1}(\Delta Y_t)^k \preceq (k!/2)c^{k-2}\mathbb{E}_{t-1}\Delta Y_t^2$.

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 (29) 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 (Rudelson, 1999; Vershynin, 2012; Gittens and Tropp, 2011; Tropp, 2015; Koltchinskii and Lounici, 2017). 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 $\widehat{\Sigma}_t \coloneqq t^{-1}\sum_{i=1}^t x_i x_i^T$. For fixed-sample estimation, when $\|x_i\|_2 \le \sqrt{b}$ a.s. for all $i \in [t]$, the analysis of Tropp (2015, section 1.6.3) implies

$$\mathbb{P}\left(\|\widehat{\Sigma}_t - \Sigma\|_{\text{op}} \ge \sqrt{\frac{2b\|\Sigma\|_{\text{op}}\log(2d/\alpha)}{t}} + \frac{4b\log(2d/\alpha)}{3t}\right) \le \alpha. \tag{30}$$

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*

$$\mathbb{P}\left(\exists t \ge 1 : \|\widehat{\Sigma}_t - \Sigma\|_{op} \ge \frac{1}{t}u\left(bt\|\Sigma\|_{op}\right)\right) \le \alpha. \tag{31}$$

For example, using the polynomial stitched bound with scale $c = 2b/3$ and $m = b\|\Sigma\|_{\text{op}}$, 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

$$\mathbb{P}\left(\exists t \ge 1 : \|\widehat{\Sigma}_t - \Sigma\|_{\text{op}} \ge k_1\sqrt{\frac{b\|\Sigma\|_{\text{op}}\ell(t)}{t}} + \frac{4bk_2\ell(t)}{3t}\right) \le \alpha, \tag{32}$$

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 (8).

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

$$\|\widehat{\Sigma}_t - \Sigma\|_{\text{op}} \lesssim \sqrt{\frac{b\log(d\log t)}{t}} + \frac{b\log(d\log t)}{t}. \tag{33}$$

Compared to the fixed-sample result (30), 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 (85).

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, $\psi_\mu(\lambda) \coloneqq A(\lambda + \theta(\mu)) - A(\theta(\mu)) - \lambda\mu$, with $\psi_\mu(\lambda) \coloneqq \infty$ if the RHS does not exist. Writing $S_t(\mu) \coloneqq \sum_{i=1}^t T(X_i) - t\mu$, 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 (1976*b*, Theorem 1) for a related result.

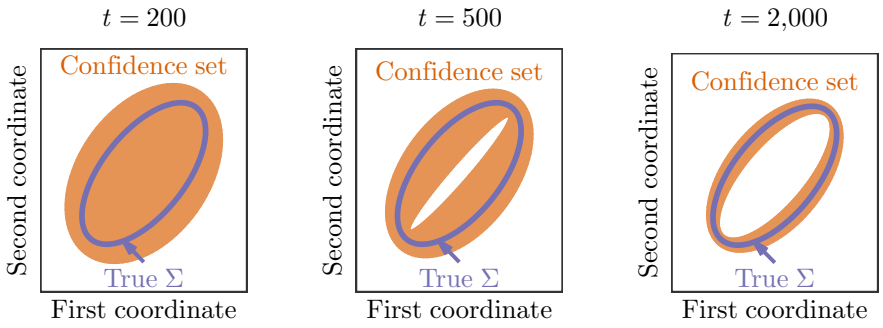

Figure 6: The matrix confidence sequence of Corollary 5 based on one simulated sequence. Observations are drawn i.i.d. taking values $\pm(\sqrt{2}\ \ \sqrt{2})^T$, $\pm(1/\sqrt{2}\ \ -1/\sqrt{2})^T$ each with probability 1/4, with covariance matrix $\Sigma = \frac{1}{4}\left(\begin{smallmatrix} 5 & 3 \\ 3 & 5 \end{smallmatrix}\right)$, 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 (85) with $s = 1.4$ matching (11), $\eta = 1.1$ and $\overline{\lambda} = 0.262$ chosen as described in Appendix E.

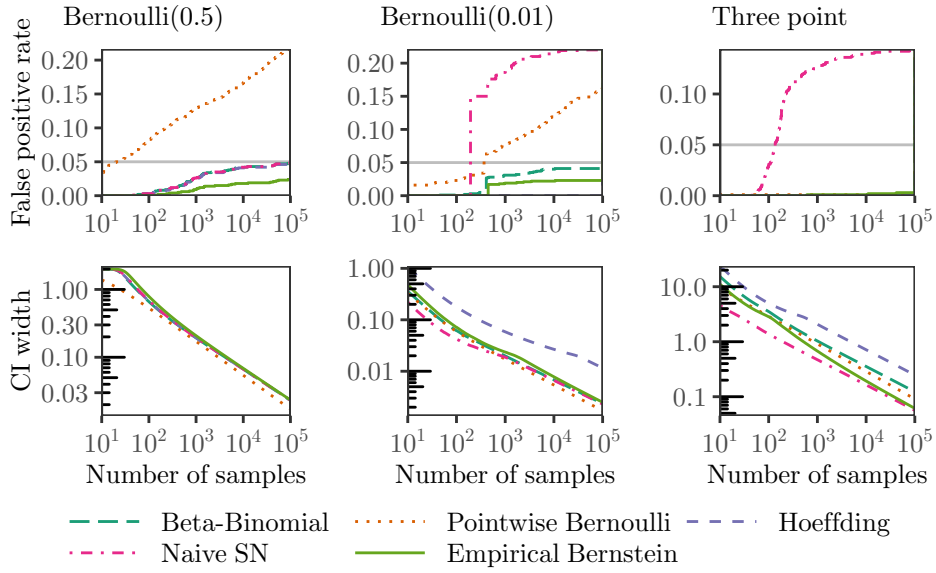

Figure 7: Summary of 1,000 simulations, each with 100,000 i.i.d. observations from the indicated distribution. Top panels show the proportion of replications in which the 95%-confidence sequence has excluded the true mean by time t. Bottom panels show the mean confidence interval width. The "three point" 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 (14), while"Beta-Binomial" uses the beta-binomial mixture (Proposition 7). "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 9. "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 $t = 500$.

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*

$$\mathrm{CI}_t \coloneqq \{\mu \in \Omega : -\tilde{u}_\mu(t) < S_t(\mu) < u_\mu(t)\}\,, \tag{34}$$

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 (Hoeffding, 1963; cf. Howard et al., 2020, Lemma 3(c)). We use a two-sided normal mixture boundary (14) 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 (Hoeffding, 1963; cf. Howard et al., 2020, 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 7, 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 $\{\mu \in [a, b] : -f_{g(\mu),h(\mu)}(V_t(\mu)) \leq \sum_{i=1}^t X_i - t\mu \leq f_{h(\mu),g(\mu)}(V_t(mu))\}$. This is more efficiently computed using the mixture supermartingale $m(S_t, V_t)$ of (57), 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 $V_t \psi_B^\star(S_t/V_t) \geq \log \alpha^{-1}$, 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 (Proposition 9). 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 (14). 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. (2017) 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.

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 (Johari et al., 2015) for sequential tests. Indeed, a slew of related

[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.

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) $\mathbb{P}\left(\bigcup_{t=1}^\infty A_t\right) \le \alpha$.

(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), based on Darling and Robbins (1967*a*) and Lai (1984), differs from that Johari et al. (2015), 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. (2016).

Always-valid p-values and tests of power one As an alternative to confidence sequences, Johari et al. (2015) 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 (Johari et al., 2015, 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* (Darling and Robbins, 1967*b*; Robbins, 1970). 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 (34) 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 (Robbins, 1970) 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) 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 $\mathbb{P}(\forall t : \theta \in \mathrm{CI}_t) \ge 1-\alpha$, then the running intersection $\widetilde{\mathrm{CI}}_t := \cap_{s\le t}\mathrm{CI}_t$ is also uniformly valid for θ, is never larger and may be much smaller. This was observed by Darling and Robbins (1967*b*), and is used in the implementation of Johari et al. (2017), for example. (In the language of sequential testing, if $(p_t)_{t=1}^\infty$ is an always-valid p-value process, then so is $(\min_{s\le t} p_s)_{t=1}^\infty$.)

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

non-stationary case, the non-intersected sequence is the more sensible one to use. The solution of Johari et al. (2017) 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 (1989) make a case for the non-intersected 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 non-intersected 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. (2020), 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 (1972, 1974) 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 (Wald, 1947, section 10.5), (Lorden and Pollak, 2005)). 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 (Siegmund and Gregory, 1980; Lai, 1997; Kulldorff et al., 2011).

The concept of *test (super)martingales* expounded by Shafer et al. (2011) 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 (1939) and discussed by Robbins (1970) and Lai (1976*b*). A recent "safe testing" framework of Grünwald et al. (2019) 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 (Pocock, 1977; O'Brien and Fleming, 1979; Lan and DeMets, 1983; Jennison and Turnbull, 2000). 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 (1978) gave an analytical treatment of the problem; numerical methods are also available for group sequential tests (Jennison and Turnbull, 2000, 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 (2014, 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 Garivier and Leonardi (2011). 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 (Lai and Siegmund, 1977, 1979; Siegmund, 1985; Whitehead and Stratton, 1983). It would be interesting to understand whether such improvements work out in nonparametric settings.

Acknowledgments

Howard thanks ONR Grant N00014-15-1-2367. Sekhon thanks ONR grants N00014-17-1-2176 and N00014-15-1-2367. Ramdas thanks NSF grant DMS1916320. We thank Boyan Duan and Ian Waudby-Smith as well as the referees/AE for useful suggestions.

References

Armitage, P., McPherson, C. K. and Rowe, B. C. (1969), 'Repeated significance tests on accumulating data', *Journal of the Royal Stat. Society, Series A* **132**(2), 235–244.

Aronow, P. M. and Middleton, J. A. (2013), 'A class of unbiased estimators of the average treatment effect in randomized experiments', *Journal of Causal Inference* **1**(1), 135–154.

Audibert, J.-Y., Munos, R. and Szepesvári, C. (2009), 'Exploration–exploitation tradeoff using variance estimates in multi-armed bandits', *Theoretical Computer Science* **410**(19), 1876–1902.

Azuma, K. (1967), 'Weighted sums of certain dependent random variables.', *Tohoku Mathematical Journal* **19**(3), 357–367.

Balsubramani, A. (2014), 'Sharp Finite-Time Iterated-Logarithm Martingale Concentration', *arXiv:1405.2639* .

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, AUAI Press, pp. 42–51.

Bercu, B., Delyon, B. and Rio, E. (2015), *Concentration Inequalities for Sums and Martingales*, Springer International Publishing, Cham.

Bercu, B. and Touati, A. (2008), 'Exponential inequalities for self-normalized martingales with applications', *The Annals of Applied Probability* **18**(5), 1848–1869.

Berman, R., Pekelis, L., Scott, A. and Van den Bulte, C. (2018), p-hacking and false discovery in A/B testing, Technical Report 3204791, SSRN.

Boucheron, S., Lugosi, G. and Massart, P. (2013), *Concentration inequalities: a nonasymptotic theory of independence*, 1st edn, Oxford University Press, Oxford.

Chernoff, H. (1952), 'A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations', *The Annals of Mathematical Statistics* **23**(4), 493–507.

Cramér, H. (1938), 'Sur un nouveau théorème-limite de la théorie des probabilités', *Actualités Scientifiques* **736**.

Darling, D. A. and Robbins, H. (1967*a*), 'Confidence Sequences for Mean, Variance, and Median', *Proceedings of the National Academy of Sciences* **58**(1), 66–68.

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

Darling, D. A. and Robbins, H. (1968), 'Some Further Remarks on Inequalities for Sample Sums', *Proceedings of the National Academy of Sciences* **60**(4), 1175–1182.

de la Peña, V. H. (1999), 'A General Class of Exponential Inequalities for Martingales and Ratios', *The Annals of Probability* **27**(1), 537–564.

de la Peña, V. H., Klass, M. J. and Lai, T. L. (2004), 'Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws', *The Annals of Probability* **32**(3), 1902–1933.

de la Peña, V. H., Klass, M. J. and Lai, T. L. (2007), 'Pseudo-maximization and self-normalized processes', *Probability Surveys* **4**, 172–192.

de la Peña, V. H., Klass, M. J. and Lai, T. L. (2009), 'Theory and applications of multivariate self-normalized processes', *Stochastic Processes and their Applications* **119**(12), 4210–4227.

de la Peña, V. H., Lai, T. L. and Shao, Q.-M. (2009), *Self-normalized processes: limit theory and statistical applications*, Springer, Berlin.

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

Durrett, R. (2017), *Probability: Theory and Examples*, 5a edn.

Efron, B. (1971), 'Forcing a sequential experiment to be balanced', *Biometrika* **58**(3), 403–417.

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

Freedman, D. A. (1975), 'On Tail Probabilities for Martingales', *The Annals of Probability* **3**(1), 100–118.

Fulks, W. (1951), 'A Generalization of Laplace's Method', *Proceedings of the American Mathematical Society* **2**(4), 613–622.

Garivier, A. (2013), Informational confidence bounds for self-normalized averages and applications, *in* '2013 IEEE Information Theory Workshop (ITW)', IEEE, pp. 1–5.

Garivier, A. and Leonardi, F. (2011), 'Context tree selection: A unifying view', *Stochastic Processes and their Applications* **121**(11), 2488–2506.

Gittens, A. and Tropp, J. A. (2011), 'Tail bounds for all eigenvalues of a sum of random matrices', *ACM Report 2014-02, Caltech* .

Grünwald, P., de Heide, R. and Koolen, W. (2019), 'Safe testing', *arXiv:1906.07801* .

Hoeffding, W. (1963), 'Probability Inequalities for Sums of Bounded Random Variables', *Journal of the American Statistical Association* **58**(301), 13–30.

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

Imbens, G. W. and Rubin, D. B. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, 1 edn, Cambridge University Press.

Jamieson, K. and Jain, L. (2018), A bandit approach to multiple testing with false discovery control, *in* 'Proceedings of the 32nd International Conference on Neural Information Processing Systems', pp. 3664–3674.

Jamieson, K., Malloy, M., Nowak, R. and Bubeck, S. (2014), lil' UCB: An Optimal Exploration Algorithm for Multi-Armed Bandits, *in* 'Proceedings of The 27th Conference on Learning Theory', Vol. 35, pp. 423–439.

Jamieson, K. and Nowak, R. (2014), Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting, *in* '48th Annual Conference on Information Sciences and Systems (CISS)', pp. 1–6.

Jennison, C. and Turnbull, B. W. (1984), 'Repeated confidence intervals for group sequential clinical trials', *Controlled Clinical Trials* **5**(1), 33–45.

Jennison, C. and Turnbull, B. W. (1989), 'Interim Analyses: The Repeated Confidence Interval Approach', *Journal of the Royal Statistical Society, Series B* **51**(3), 305–361.

Jennison, C. and Turnbull, B. W. (2000), *Group sequential methods with applications to clinical trials*, Chapman & Hall/CRC, Boca Raton.

Johari, R., Koomen, P., Pekelis, L. and Walsh, D. (2017), Peeking at A/B Tests: Why it matters, and what to do about it, ACM Press, pp. 1517–1525.

Johari, R., Pekelis, L. and Walsh, D. J. (2015), 'Always valid inference: Bringing sequential analysis to A/B testing', *arXiv preprint arXiv:1512.04922* .

Jorgensen, B. (1997), *The Theory of Dispersion Models*, CRC Press.

Kaufmann, E., Cappé, O. and Garivier, A. (2016), 'On the complexity of best-arm identification in multi-armed bandit models', *The Journal of Machine Learning Research* **17**(1), 1–42.

Kaufmann, E. and Koolen, W. (2018), 'Mixture martingales revisited with applications to sequential tests and confidence intervals', *arXiv:1811.11419* .

Koltchinskii, V. and Lounici, K. (2017), 'Concentration inequalities and moment bounds for sample covariance operators', *Bernoulli* **23**(1), 110–133.

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 Analysis* **30**(1), 58–78.

Lai, T. L. (1976*a*), 'Boundary Crossing Probabilities for Sample Sums and Confidence Sequences', *The Annals of Probability* **4**(2), 299–312.

Lai, T. L. (1976*b*), 'On Confidence Sequences', *The Annals of Statistics* **4**(2), 265–280.

Lai, T. L. (1984), 'Incorporating scientific, ethical and economic considerations into the design of clinical trials in the pharmaceutical industry: a sequential approach', *Communications in Statistics - Theory and Methods* **13**(19), 2355–2368.

Lai, T. L. (1997), 'On optimal stopping problems in sequential hypothesis testing', *Statistica Sinica* **7**(1), 33–51.

Lai, T. L. and Siegmund, D. (1977), 'A Nonlinear Renewal Theory with Applications to Sequential Analysis I', *The Annals of Statistics* **5**(5), 946–954.

Lai, T. L. and Siegmund, D. (1979), 'A Nonlinear Renewal Theory with Applications to Sequential Analysis II', *The Annals of Statistics* **7**(1), 60–76.

Lan, K. K. G. and DeMets, D. L. (1983), 'Discrete Sequential Boundaries for Clinical Trials', *Biometrika* **70**(3), 659–663.

Lorden, G. and Pollak, M. (2005), 'Nonanticipating estimation applied to sequential analysis and changepoint detection', *The Annals of Statistics* **33**(3), 1422–1454.

Malek, A., Katariya, S., Chow, Y. and Ghavamzadeh, M. (2017), Sequential multiple hypothesis testing with Type I error control, *in* 'Artificial Intelligence and Statistics', pp. 1468–1476.

Maurer, A. and Pontil, M. (2009), Empirical Bernstein bounds and sample variance penalization, *in* 'Proceedings of the Conference on Learning Theory'.

McDiarmid, C. (1998), Concentration, *in* M. Habib, C. McDiarmid, J. Ramirez-Alfonsin and B. Reed, eds, 'Probabilistic Methods for Algorithmic Discrete Mathematics', Springer, New York, pp. 195–248.

Morters, P. and Peres, Y. (2010), *Brownian Motion*, Cambridge University Press, Cambridge.

Neyman, J. (1923/1990), 'On the Application of Probability Theory to Agricultural Experiments, Essay on Principles, Section 9', *Statistical Science* **5**(4), 465–480.

O'Brien, P. C. and Fleming, T. R. (1979), 'A Multiple Testing Procedure for Clinical Trials', *Biometrics* **35**(3), 549–556.

Pinelis, I. (1992), An Approach to Inequalities for the Distributions of Infinite-Dimensional Martingales, *in* 'Probability in Banach Spaces, 8: Proceedings of the Eighth International Conference', Birkhäuser, Boston, MA, pp. 128–134.

Pinelis, I. (1994), 'Optimum Bounds for the Distributions of Martingales in Banach Spaces', *The Annals of Probability* **22**(4), 1679–1706.

Pocock, S. J. (1977), 'Group Sequential Methods in the Design and Analysis of Clinical Trials', *Biometrika* **64**(2), 191–199.

Raginsky, M., Sason, I. et al. (2013), 'Concentration of measure inequalities in information theory, communications, and coding', *Foundations and Trends in Communications and Information Theory* **10**(1-2), 1–246.

Robbins, H. (1970), 'Statistical Methods Related to the Law of the Iterated Logarithm', *The Annals of Mathematical Statistics* **41**(5), 1397–1409.

Robbins, H. and Siegmund, D. (1968), Iterated logarithm inequalities and related statistical procedures, *in* 'Mathematics of the Decision Sciences, Part II', American Mathematical Society, Providence, pp. 267–279.

Robbins, H. and Siegmund, D. (1969), 'Probability Distributions Related to the Law of the Iterated Logarithm', *Proc. of the National Academy of Sciences* **62**(1), 11–13.

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

Robbins, H. and Siegmund, D. (1972), A class of stopping rules for testing parametric hypotheses, The Regents of the University of California.

Robbins, H. and Siegmund, D. (1974), 'The Expected Sample Size of Some Tests of Power One', *The Annals of Statistics* **2**(3), 415–436.

Rubin, D. B. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies.', *Journal of educational Psychology* **66**(5), 688.

Rudelson, M. (1999), 'Random Vectors in the Isotropic Position', *Journal of Functional Analysis* **164**(1), 60–72.

Shafer, G., Shen, A., Vereshchagin, N. and Vovk, V. (2011), 'Test Martingales, Bayes Factors and p-Values', *Statistical Science* **26**(1), 84–101.

Siegmund, D. (1978), 'Estimation Following Sequential Tests', *Biometrika* **65**(2), 341.

Siegmund, D. (1985), *Sequential Analysis*, Springer New York, New York, NY.

Siegmund, D. and Gregory, P. (1980), 'A Sequential Clinical Trial for Testing $p_1 = p_2$', *The Annals of Statistics* **8**(6), 1219–1228.

Stout, W. F. (1970), 'The Hartman-Wintner Law of the Iterated Logarithm for Martingales', *Annals of Mathematical Statistics* **41**(6), 2158–2160.

Tropp, J. A. (2011), 'Freedman's inequality for matrix martingales', *Electronic Communications in Probability* **16**, 262–270.

Tropp, J. A. (2012), 'User-friendly tail bounds for sums of random matrices', *Foundations of Computational Mathematics* **12**(4), 389–434.

Tropp, J. A. (2015), 'An Introduction to Matrix Concentration Inequalities', *Foundations and Trends in Machine Learning* **8**(1-2), 1–230.

van de Geer, S. (1995), 'Exponential Inequalities for Martingales, with Application to Maximum Likelihood Estimation for Counting Processes', *The Annals of Statistics* **23**(5), 1779–1801.

Vershynin, R. (2012), Introduction to the non-asymptotic analysis of random matrices, *in* 'Compressed Sensing: Theory and Applications', Cambridge University Press.

Ville, J. (1939), *Étude Critique de la Notion de Collectif.*, Gauthier-Villars, Paris.

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

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

Whitehead, J. and Stratton, I. (1983), 'Group Sequential Clinical Trials with Triangular Continuation Regions', *Biometrics* **39**(1), 227–236.

Widder, D. V. (1942), *Laplace Transform*, Princeton University Press, Princeton.

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* '31st Conference on Neural Information Processing Systems'.

Zhao, S., Zhou, E., Sabharwal, A. and Ermon, S. (2016), Adaptive concentration inequalities for sequential decision problems, *in* '30th Conference on Neural Information Processing Systems'.

A Proofs of main results

In this section we give proofs of our main results along with selected discussion of and intuition for proof techniques.

A.1 Proof of Theorem 1

The idea behind Theorem 1 is to divide intrinsic time into geometrically spaced epochs, $\eta^k \le V_t < \eta^{k+1}$ for some $\eta > 1$. We construct a linear boundary within each epoch using Lemma 1 and take a union bound over crossing events of the different boundaries. The resulting, piecewise-linear boundary may then be upper bounded by a smooth, concave function. Figure 3 illustrates the construction.

As discussed in Section 3.1, the function h determines the nominal crossing probability $\alpha/h(k)$ allocated to the k^{th} epoch, and we have mentioned the choices $h(k) = \eta^{sk}/(1-\eta^{-s})$ and $h(k) = (k+1)^s\zeta(s)$. One may substitute a series converging yet more slowly; for example, $h(k) \propto (k+2)\log^s(k+2)$ for $s > 1$ yields

$$\log h(\log_\eta V_t) = \log\log_\eta(\eta^2 V_t) + s\log\log\log_\eta(\eta^2 V_t) + \log\left(\frac{\log^{1-s}(3/2)}{s-1}\right), \tag{35}$$

matching related analysis in Darling and Robbins (1967*b*), Robbins and Siegmund (1969), Robbins (1970), and Balsubramani (2014). In practice, the bound (35) appears to behave like bound (10) with worse constants. However, the fact that the stitching approach can recover key theoretical results like these gives some indication of its power.

Proof of Theorem 1. We prove the result in the case $m = 1$ for simplicity. The general result may be obtained by considering $S_t/\sqrt{m}$ in place of S_t, V_t/m in place of V_t, and $c/\sqrt{m}$ in place of c. See Appendix F for details.

We first compute $\psi_G^{-1}(u)$ by taking the positive solution to the quadratic equation given by $\psi_G(\lambda) = u$, yielding

$$\psi_G^{-1}(u) = -cu \pm \sqrt{c^2u^2 + 2u} = \frac{2}{c + \sqrt{c^2 + 2/u}}, \tag{36}$$

where we have used the identity $\sqrt{1+x} - 1 = \frac{x}{\sqrt{1+x}+1}$. Let

$$K(u) := \frac{\sqrt{2u}}{\psi_G^{-1}(u)} = \sqrt{1 + \frac{c^2u}{2}} + c\sqrt{\frac{u}{2}}. \tag{37}$$

$K(u)$ will appear below. Now we start from the line-crossing inequality of Lemma 1: reparametrizing $r = \log \alpha^{-1}$, we have for any $r > 0, \lambda > 0$

$$\mathbb{P}\left(\exists t \geq 1 : S_t \geq \underbrace{\frac{r + \psi_G(\lambda) V_t}{\lambda}}_{g_{\lambda,r}(V_t)}\right) \leq l_0 e^{-r}. \tag{38}$$

We divide intrinsic time into epochs $\eta^k \leq V_t < \eta^{k+1}$ for each $k = 0, 1, \ldots$, and we will construct a linear boundary over each epoch by carefully choosing values for λ_k and r_k and using the probability bound (38). We choose λ_k so that the "standardized" boundary takes equal values at both endpoints of the epoch: $g_{\lambda_k,r_k}(\eta^k)/\eta^{k/2} = g_{\lambda_k,r_k}(\eta^{k+1})/\eta^{(k+1)/2}$. This equation is solved by $\lambda_k = \psi_G^{-1}(r_k/\eta^{k+1/2})$, which yields, after some algebra,

$$g_{\lambda_k,r_k}(v) = K\left(\frac{r_k}{\eta^{k+1/2}}\right)\left[\sqrt{\frac{\eta^{k+1/2}}{v}} + \sqrt{\frac{v}{\eta^{k+1/2}}}\right]\sqrt{\frac{r_k v}{2}} \tag{39}$$

Our goal, after choosing r_k below, is to upper bound this expression by a function of v alone, independent of k. Noting that the term in square brackets in (39) reaches its maximum over the k^{th} epoch at the endpoints, $v = \eta^k$ and $v = \eta^{k+1}$, and substituting the expression (37) for $K(u)$, we have

$$g_{\lambda_k,r_k}(v) \leq \left(\sqrt{1 + \frac{c^2 r_k}{2\eta^{k+1/2}}} + c\sqrt{\frac{r_k}{2\eta^{k+1/2}}}\right)\frac{\eta^{1/4} + \eta^{-1/4}}{\sqrt{2}}\sqrt{r_k v}, \quad \text{for all } \eta^k \leq v < \eta^{k+1}. \tag{40}$$

The inequality $\eta^{k+1/2} \geq v/\sqrt{\eta}$ yields

$$g_{\lambda_k,r_k}(v) \leq \frac{\eta^{1/4} + \eta^{-1/4}}{\sqrt{2}}\left(\sqrt{r_k v + \frac{\sqrt{\eta} c^2 r_k^2}{2}} + c\frac{\eta^{1/4} r_k}{\sqrt{2}}\right) \tag{41}$$

$$= \sqrt{k_1^2 r_k v + k_2^2 c^2 r_k^2} + c k_2 r_k, \quad \text{for all } \eta^k \leq v < \eta^{k+1}, \tag{42}$$

using the definition (8) of k_1 and k_2. Now let $r_k = \log(l_0 h(k)/\alpha)$, which we choose to ensure total error probability will be bounded by α via a union bound. Note that h is nondecreasing and $k \leq \log_\eta V_t$ over the epoch, so that $r_k \leq \ell(v)$ over the epoch, recalling the definition (8) of $\ell(v)$. We conclude

$$g_{\lambda_k,r_k}(v) \leq \sqrt{k_1^2 v \ell(v) + k_2^2 c^2 \ell^2(v)} + c k_2 \ell(v) = \mathcal{S}_\alpha(v), \tag{43}$$

for all $\eta^k \leq v < \eta^{k+1}$. This final expression no longer depends on k, showing that the final boundary $\mathcal{S}_\alpha(v)$ majorizes the corresponding linear boundary $g_{\lambda_k,r_k}(v)$ over each epoch $\eta^k \leq v < \eta^{k+1}$ for $k = 0, 1, \ldots$. Hence

$$\mathcal{S}_\alpha(v) \geq \min_{k \geq 0} g_{\lambda_k,r_k}(v) \quad \text{for all } v \geq 1. \tag{44}$$

But the first linear boundary $g_{\lambda_0,t_0}(v)$ passes through $\mathcal{S}_\alpha(1)$ and has positive slope, which implies

$$\mathcal{S}_\alpha(1 \vee v) \geq \min_{k \geq 0} g_{\lambda_k,r_k}(v) \quad \text{for all } v > 0. \tag{45}$$

Now taking a union bound over the probability bounds given by (38) for $k = 0, 1, \ldots$, we have

$$\mathbb{P}\left(\exists t \geq 1 : S_t \geq \min_{k \geq 0} g_{\lambda_k,r_k}(V_t)\right) \leq l_0 \sum_{k=0}^{\infty} e^{-r_k} = \alpha \sum_{k=0}^{\infty} \frac{1}{h(k)} \leq \alpha. \tag{46}$$

Combining (46) with (45) proves that $v \mapsto \mathcal{S}_\alpha(1 \vee v)$ is a sub-gamma uniform boundary with crossing probability α.

For the second statement (9), we simply restrict the union bound to epochs $k \geq \lfloor \log_\eta V_t \rfloor$, which restricts the sum in (46) accordingly. □

We have given a stitched bound which is constant for $v < m$, but inspection of the proof shows that one may improve the bound to be linear with positive slope on $v < m$, by extending the linear bound over the first epoch to cover all $v > 0$. This seems of limited utility for theoretical work, and we recommend other bounds over the stitched bound for practice, so we do not pursue this point further.

The idea of taking a union bound over geometrically spaced epochs is standard in the proof of the classical law of the iterated logarithm (Durrett, 2017, Theorem 8.5.1). The idea has been extended to finite-time bounds by Darling and Robbins (1967*b*), Jamieson et al. (2014), Kaufmann et al. (2016), and Zhao et al. (2016), usually when the observations are independent and sub-Gaussian; the technique is sometimes called "peeling". Of course, Theorem 1 generalizes these constructions much beyond the independent sub-Gaussian case, but it also achieves tighter constants for the sub-Gaussian setting. Here, we briefly discuss how the improved constants arise.

Both Jamieson et al. (2014) and Zhao et al. (2016) construct a constant boundary rather than a linear increasing boundary over each epoch. They apply Doob's maximal inequality for submartingales (Durrett, 2017, Theorem 4.4.2), as in Hoeffding (1963, eq. 2.17), to obtain boundaries similar to that of Freedman (1975). As illustrated in Howard et al. (2020, Figure 2), the linear bounds from Lemma 1 are stronger than corresponding Freedman-style bounds, and the additional flexibility yields tighter constants.

Both Darling and Robbins (1967*b*) and Kaufmann et al. (2016) use linear boundaries within each epoch analogous to those of Lemma 1. Both methods share a great deal in common with ours, and Darling and Robbins give consideration to general cumulant-generating functions. Recall from Lemma 1 that such linear boundaries may be chosen to optimize for some fixed time $V_t = m$. Our method chooses the linear boundary within each epoch to be optimal at the geometric center of the epoch, i.e., at $V_t = \eta^{k+1/2}$, so that at both epoch endpoints the boundary will be equally "loose", that is, equal multiples of $\sqrt{V_t}$. Darling and Robbins choose the boundaries to be tangent at the start of the epoch, hence their boundary is looser than ours at the end of the epoch. Kaufmann et al. choose the boundary as we do, but appear to incur more looseness in the subsequent inequalities used to construct a smooth upper bound.

A.2 Proof of Corollary 1

Fix any $\epsilon > 0$ and choose $a > 0$ small enough that $\psi(\lambda) \leq (1+\epsilon)\lambda^2/2$ for all $\lambda \in (0, a)$. Using the fact that $\psi_{G,c}(\lambda) \geq \lambda^2/2$ for $c \geq 0$, we have $\psi(\lambda) \leq (1+\epsilon)\psi_{G,1/a}(\lambda)$ for all $\lambda \in (0, a)$, so that (S_t) is sub-gamma with scale $c = 1/a$ and variance process $((1+\epsilon)V_t)$. Now Theorem 1 shows that

$$\mathbb{P}\left(\sup_t V_t = \infty \text{ and } S_t \geq u((1+\epsilon)V_t) \text{ infinitely often}\right) = 0, \tag{47}$$

where we may choose $u(v) \sim \sqrt{2(1+\epsilon)v \log\log v}$ (see (10) and discussion thereafter), so that $u((1+\epsilon)v) \sim \sqrt{2(1+\epsilon)^2 v \log\log v}$. It follows that

$$\limsup_{t\to\infty} \frac{S_t}{\sqrt{2(1+\epsilon)^2 V_t \log\log V_t}} \leq 1 \quad \text{on} \quad \left\{\sup_t V_t = \infty\right\}. \tag{48}$$

As $\epsilon > 0$ was arbitrary, we are done. □

A.3 Conjugate mixture proofs

Proof of Lemma 2. Assume (S_t) is sub-ψ with variance process (V_t), so that, for each $\lambda \in [0, \lambda_{\max})$, we have $\exp\{\lambda S_t - \psi(\lambda)V_t\} \leq L_t(\lambda)$ where $(L_t(\lambda))_{t=0}^{\infty}$ is a nonnegative supermartingale. We will show that $M_t \coloneqq \int L_t(\lambda)\,\mathrm{d}F(\lambda)$ is a supermartingale with respect to $(\mathcal{F}_t)$.

Formally, for this proof, we augment the underlying probability space with the random variable λ having distribution F over the Borel σ-field on $\mathbb{R}$, independent of everything else. For each t, we require L_t to be a random variable on this product space, i.e., it must be product measurable. Now Definition 1 stipulates that $L_t \in \sigma(\lambda, \mathcal{F}_t)$ and $\mathbb{E}(L_t \mid \lambda, \mathcal{F}_{t-1}) \leq L_{t-1}$ for each $t \geq 1$, and additionally, $\mathbb{E}(L_0 \mid \lambda) \leq l_0$ a.s. In other words, (L_t) is a supermartingale with respect to the filtration given by $\mathcal{G}_t \coloneqq \sigma(\lambda, \mathcal{F}_t)$ on this augmented space. Finally, we have $M_t = \mathbb{E}(L_t \mid \mathcal{F}_t)$. These facts follow directly from the definition and properties of conditional expectation.

We claim that (M_t) is a supermartingale with respect to $(\mathcal{F}_t)$ on this augmented space. Indeed,

$$\mathbb{E}\,(\,M_t \mid \mathcal{F}_{t-1}) = \mathbb{E}\,(\,\mathbb{E}\,(\,L_t \mid \mathcal{F}_t) \mid \mathcal{F}_{t-1}) = \mathbb{E}\,(\,\mathbb{E}\,(\,L_t \mid \lambda, \mathcal{F}_{t-1}) \mid \mathcal{F}_{t-1}) \leq \mathbb{E}\,(\,L_{t-1} \mid \mathcal{F}_{t-1}) \tag{49}$$

by the supermartingale property, and this last expression is equal to M_{t-1}. Furthermore, $\mathbb{E}M_0 = \mathbb{E}\mathbb{E}\,(\,L_0 \mid \lambda) \leq l_0$ since $\mathbb{E}\,(\,L_0 \mid \lambda) \leq l_0$ a.s., hence $\mathbb{E}|M_t| = \mathbb{E}M_t \leq l_0$ for all t.

Now Definition 1 and Ville's maximal inequality for nonnegative supermartingales (Durrett, 2017, exercise 4.8.2) yield

$$\mathbb{P}\left(\exists t \geq 1 : \int \exp\{\lambda S_t - \psi(\lambda) V_t\}\,\mathrm{d}F(\lambda) \geq \frac{l_0}{\alpha}\right) \leq \mathbb{P}\left(\exists t \geq 1 : M_t \geq \frac{l_0}{\alpha}\right) \leq \alpha. \tag{50}$$

In other words, $\mathbb{P}(\exists t \geq 1 : S_t \geq \mathcal{M}_\alpha(V_t)) \leq \alpha$ by the definition of $\mathcal{M}_\alpha$, which is the desired conclusion. □

In the sub-Gaussian case, the following boundary is well-known (Robbins, 1970, example 2).

Proposition 5 (Two-sided normal mixture)**.** *Suppose both (S_t) and $(-S_t)$ are sub-Gaussian with variance process (V_t). Fix $\alpha \in (0,1)$ and $\rho > 0$, and define*

$$u(v) := \sqrt{(v+\rho)\log\left(\frac{l_0^2(v+\rho)}{\alpha^2\rho}\right)}. \tag{51}$$

Then $\mathbb{P}(\forall t \geq 1 : |S_t| < u(V_t)) \geq 1 - \alpha$.

We have included the bound in Figures 4 and 9; although its $\mathcal{O}(\sqrt{V_t \log V_t})$ rate of growth is worse than the finite LIL discrete mixture bound, it can achieve tighter control over about three orders of magnitude of intrinsic time. This makes the normal mixture preferable in many practical situations when a sub-Gaussian assumption applies. When only a one-sided sub-Gaussian assumption holds, the normal mixture still yields a sub-Gaussian uniform boundary.

Proposition 6 (One-sided normal mixture)**.** *For any $\alpha \in (0,1)$ and $\rho > 0$, the boundary*

$$\mathrm{NM}_\alpha(v) = \sup\left\{s \in \mathbb{R} : \sqrt{\frac{4\rho}{v+\rho}}\exp\left\{\frac{s^2}{2(v+\rho)}\right\}\Phi\left(\frac{s}{\sqrt{v+\rho}}\right) < \frac{l_0}{\alpha}\right\}. \tag{52}$$

is a sub-Gaussian uniform boundary with crossing probability α. Furthermore, we have the following closed-form upper bound:

$$\mathrm{NM}_\alpha(v) \leq \widetilde{\mathrm{NM}}_\alpha(v) := \sqrt{2(v+\rho)\log\left(\frac{l_0}{2\alpha}\sqrt{\frac{v+\rho}{\rho}} + 1\right)}. \tag{53}$$

The boundary NM_α is easily evaluated to high precision by numerical root-finding, and the closed-form approximation is excellent: numerical calculations indicate that $\widetilde{\mathrm{NM}}_{0.025}(v)/\mathrm{NM}_{0.025}(v) < 1.007$ uniformly when $\rho = 1$, for example.

Proof of Proposition 6. To obtain the explicit upper bound $\widetilde{\mathrm{NM}}_\alpha$ in (53) from the exact boundary (52), we use the inequality $1 - \Phi(x) \leq e^{-x^2/2}$ for $x > 0$, which follows from a standard Cramér-Chernoff bound. This implies

$$\sqrt{\frac{4\rho}{v+\rho}}\exp\left\{\frac{s^2}{2(v+\rho)}\right\}\Phi\left(\frac{s}{\sqrt{v+\rho}}\right) \geq \sqrt{\frac{4\rho}{v+\rho}}\left[\exp\left\{\frac{s^2}{2(v+\rho)}\right\} - 1\right], \quad \text{for } s > 0. \tag{54}$$

We set the RHS equal to l_0/α and solve to conclude

$$\mathrm{NM}_\alpha(v) \leq \sqrt{2(v+\rho)\log\left(\frac{l_0}{2\alpha}\sqrt{\frac{v+\rho}{\rho}} + 1\right)} = \widetilde{\mathrm{NM}}_\alpha(v), \tag{55}$$

so long as $\mathrm{NM}_\alpha(v) > 0$. But we are guaranteed that $\mathrm{NM}_\alpha(v) > 0$, because the LHS of the inequality in (52) is increasing in s on $s \geq 0$ and no larger than one when $s = 0$, while the RHS $l_0/\alpha \geq 1$.

The fact that NM_α is a sub-Gaussian uniform boundary follows directly from Lemma 2, and therefore $\widetilde{\mathrm{NM}}_\alpha$ is as well. □

When a sub-Bernoulli condition holds, as with bounded observations, the following beta-binomial boundary is tighter than the normal mixture. Simpler versions of this boundary have long been studied for i.i.d. Bernoulli sampling (Ville, 1939; Robbins, 1970; Lai, 1976*b*; Shafer et al., 2011). Below, $B_x(a,b) = \int_0^x p^{a-1}(1-p)^{b-1}\,\mathrm{d}p$ denotes the incomplete Beta function, whose implementation is available in statistical software packages; B_1 is the ordinary Beta function.

Proposition 7 (Two-sided beta-binomial mixture)**.** *Suppose (S_t) is sub-Bernoulli with variance process (V_t) and range parameters g, h, while $(-S_t)$ is sub-Bernoulli with variance process (V_t) and range parameters h, g. Fix any $\rho > gh$, let $r = \rho - gh$, and define*

$$f_{g,h}(v) \;\coloneqq\; \sup\left\{ s \in \left[0, \frac{r+v}{g}\right) : m_{g,h}(s,v) < \frac{l_0}{\alpha} \right\}, \tag{56}$$

$$\text{where} \quad m_{g,h}(s,v) \;\coloneqq\; \frac{(g+h)^{v/gh}}{\left[g^{v/h+s} h^{v/g-s}\right]^{1/(g+h)}} \cdot \frac{B_1\left(\frac{r+v-gs}{g(g+h)}, \frac{r+v+hs}{h(g+h)}\right)}{B_1\left(\frac{r}{g(g+h)}, \frac{r}{h(g+h)}\right)}. \tag{57}$$

Then $\mathbb{P}(\forall t \geq 1 : -f_{g,h}(V_t) < S_t < f_{h,g}(V_t)) \geq 1 - \alpha$.

As with the normal mixture, we have a one-sided variant as well.

Proposition 8 (One-sided beta-binomial mixture)**.** *Fix any $g, h > 0$, $\alpha \in (0,1)$, and $\rho > gh$. Let $r = \rho - gh$ and define*

$$f_{g,h}(v) \;\coloneqq\; \sup\left\{ s \in \left[0, \frac{r+v}{g}\right) : m_{g,h}(s,v) < \frac{l_0}{\alpha} \right\}, \tag{58}$$

$$\text{where} \quad m_{g,h}(s,v) \;\coloneqq\; \frac{(g+h)^{v/gh}}{\left[g^{v/h+s} h^{v/g-s}\right]^{1/(g+h)}} \cdot \frac{B_{h/(g+h)}\left(\frac{r+v-gs}{g(g+h)}, \frac{r+v+hs}{h(g+h)}\right)}{B_{h/(g+h)}\left(\frac{r}{g(g+h)}, \frac{r}{h(g+h)}\right)}. \tag{59}$$

Then $f_{g,h}$ is a sub-Bernoulli uniform boundary with crossing probability α and range parameters g, h.

In the sub-Bernoulli case, we first rewrite the exponential process $\exp\{\lambda S_t - \psi_B(\lambda) V_t\}$ in terms of the transformed parameter $p = [1 + (h/g)e^{-\lambda}]^{-1}$. This is motivated by the transform from the canonical parameter to the mean parameter of a Bernoulli family, but keep in mind that we make no parametric assumption here, these are merely analytical manipulations. Then a truncated Beta distribution on $p \in [g/(g+h), 1]$ yields the one-sided beta-binomial uniform boundary, while an untruncated mixture yields the two-sided boundary.

Proof of Propositions 7 and 8. For simplicity of notation, we will assume here that the problem has been scaled so that $g + h = 1$, e.g., by replacing X_t with $X_t/(g+h)$. Using the sub-Bernoulli ψ function $\psi_B(\lambda) = \frac{1}{gh}\log\left(ge^{h\lambda} + he^{-g\lambda}\right)$, the exponential integrand in our mixture is

$$\exp\left\{\lambda s - \frac{v}{gh}\log\left(ge^{h\lambda} + he^{-g\lambda}\right)\right\} = \frac{p^{v/h+s}(1-p)^{v/g-s}}{g^{v/h+s} h^{v/g-s}}, \tag{60}$$

after substituting the one-to-one transformation

$$p = p(\lambda) \coloneqq \frac{ge^{h\lambda}}{ge^{h\lambda} + he^{-g\lambda}}, \quad \text{so that} \quad \lambda = \log\left(\frac{ph}{(1-p)g}\right), \tag{61}$$

followed by some algebra. We wish to integrate against a Beta mixture density on p with parameters r/h and r/g, which has mean $p = g$, corresponding to $\lambda = 0$. For Proposition 8, we must also truncate to $\lambda \geq 0$, i.e., to $p \geq g$. The appropriately normalized mixture integral is then

$$\frac{1}{g^{v/h+s} h^{v/g-s}} \cdot \frac{\int_g^1 p^{v/h+s+r/h-1}(1-p)^{v/g-s+r/g-1}\,\mathrm{d}p}{\int_g^1 p^{r/h-1}(1-p)^{r/g-1}\,\mathrm{d}p} = \frac{1}{g^{v/h+s} h^{v/g-s}} \cdot \frac{B_h\left(\frac{r+v}{g} - s, \frac{r+v}{h} + s\right)}{B_h\left(\frac{r}{g}, \frac{r}{h}\right)}, \tag{62}$$

using the fact that $B_x(a,b) = \int_0^x p^{a-1}(1-p)^{b-1}\,\mathrm{d}p = \int_{1-x}^1 p^{b-1}(1-p)^{a-1}\,\mathrm{d}p$. This gives the closed-form mixture (59). (To obtain the formula for general $g+h \neq 1$, substitute $g/(g+h)$ for g, $h/(g+h)$ or h, $s/(g+h)$ for s, $v/(g+h)^2$ for v, and $r/(g+h)^2$ for r.)

The proof of Proposition 7 is nearly identical, but we integrate over the full Beta mixture rather than truncating.

To verify that our choice of r ensures that λ has approximate precision ρ under the full (not truncated) mixture distribution, we use the delta method to calculate the approximate variance of λ for large r based on the variance of p under the full Beta mixture:

$$\operatorname{Var}\lambda \approx \left[\left(\frac{1}{p(1-p)}\right)^2\right]_{p=g} \cdot \frac{gh}{\frac{r}{gh}+1} = \frac{1}{r+gh}. \tag{63}$$

Setting this equal to $1/\rho$ yields $r = \rho - gh$ as desired. □

When tails are heavier than Gaussian, the normal mixture boundary is not applicable. However, the following sub-exponential mixture boundary, based on a gamma mixing density, is universally applicable, as described in Proposition 1. Like the normal mixture, the gamma-exponential mixture is unimprovable as described in Section 3.6. Below we make use of the regularized lower incomplete gamma function $\gamma(a,x) \coloneqq (\int_0^x u^{a-1}e^{-u}\,\mathrm{d}u)/\Gamma(a)$, available in standard statistical software packages.

Proposition 9 (Gamma-exponential mixture)**.** *Fix $c > 0, \rho > 0$ and define*

$$\mathrm{GE}_\alpha(v) \;\coloneqq\; \sup\left\{s \geq 0 : m(s,v) < \frac{l_0}{\alpha}\right\}, \tag{64}$$

$$\textit{where}\quad m(s,v) \;\coloneqq\; \frac{\left(\frac{\rho}{c^2}\right)^{\frac{\rho}{c^2}}}{\Gamma\left(\frac{\rho}{c^2}\right)\gamma\left(\frac{\rho}{c^2},\frac{\rho}{c^2}\right)} \frac{\Gamma\left(\frac{v+\rho}{c^2}\right)\gamma\left(\frac{v+\rho}{c^2},\frac{cs+v+\rho}{c^2}\right)}{\left(\frac{cs+v+\rho}{c^2}\right)^{\frac{v+\rho}{c^2}}} \exp\left\{\frac{cs+v}{c^2}\right\}. \tag{65}$$

Then GE_α is a sub-exponential uniform boundary with crossing probability α for scale c.

The gamma-exponential mixture is the result of evaluating the mixture integral in (13) with mixing density

$$\frac{\mathrm{d}F}{\mathrm{d}\lambda} = \frac{1}{\gamma(\rho/c^2,\rho/c^2)}\frac{(\rho/c)^{\rho/c^2}}{\Gamma(\rho/c^2)}(c^{-1}-\lambda)^{\rho/c^2-1}e^{-\rho(c^{-1}-\lambda)/c}. \tag{66}$$

This is a gamma distribution with shape ρ/c^2 and scale ρ/c applied to the transformed parameter $u = c^{-1}-\lambda$, truncated to the support $[0, c^{-1}]$. The distribution has mean zero and variance equal to $1/\rho$, making it comparable to the normal mixture distribution used above. As $\rho \to \infty$, the gamma mixture distribution converges to a normal distribution and concentrates about $\lambda = 0$, the regime in which $\psi_E(\lambda) \sim \psi_N(\lambda)$, which gives some intuition for why the gamma-exponential mixture recovers the normal mixture when $\rho \gg c^2$.

Proof of Proposition 9. We need only show that

$$m(s,v) = \int_0^{1/c} \exp\left\{\lambda s - \psi_E(\lambda)v\right\} f(\lambda)\,\mathrm{d}\lambda, \tag{67}$$

$$\text{where } f(\lambda) = \frac{1}{\gamma(\rho/c^2,\rho/c^2)}\frac{(\rho/c)^{\rho/c^2}}{\Gamma(\rho/c^2)}(c^{-1}-\lambda)^{\rho/c^2-1}e^{-\rho(c^{-1}-\lambda)/c}. \tag{68}$$

Then the fact that GM_α is a sub-exponential uniform boundary follows as a special case of Lemma 2.

Proving (67) is an exercise in calculus. Substituting the definition of ψ_E and removing common terms, it suffices to show that

$$c^{-\rho/c^2}\frac{\Gamma\left(\frac{v+\rho}{c^2}\right)\gamma\left(\frac{v+\rho}{c^2},\frac{cs+v+\rho}{c^2}\right)}{\left(\frac{cs+v+\rho}{c^2}\right)^{\frac{v+\rho}{c^2}}}e^{(cs+v)/c^2} = \int_0^{1/c}(1-c\lambda)^{v/c^2}e^{\lambda(s+v/c)}(c^{-1}-\lambda)^{\rho/c^2-1}e^{-\rho(c^{-1}-\lambda)/c}\,\mathrm{d}\lambda. \tag{69}$$

After change of variables $u = \left(\frac{cs+v+\rho}{c}\right)(c^{-1} - \lambda)$, the right-hand side is equal to

$$\left(\frac{cs+v+\rho}{c}\right)^{-\frac{v+\rho}{c^2}} c^{v/c^2} e^{(cs+v)/c^2} \int_0^{(cs+v+\rho)/c^2} u^{(v+\rho)/c^2-1} e^{-u}\,\mathrm{d}u. \tag{70}$$

Now the definition of the regularized lower incomplete gamma function and a bit of algebra finishes the argument. □

A similar mixture boundary holds in the sub-Poisson case, making use of the regularized upper incomplete gamma function $\bar{\gamma}(a,x) \coloneqq (\int_x^\infty u^{a-1}e^{-u}\,\mathrm{d}u)/\Gamma(a)$.

Proposition 10 (Gamma-Poisson mixture)**.** *Fix $c > 0, \rho > 0$ and define*

$$\mathrm{GP}_\alpha(v) \;\coloneqq\; \sup\left\{s \geq 0 : m(s,v) < \frac{l_0}{\alpha}\right\}, \tag{71}$$

$$\textit{where}\quad m(s,v) \;\coloneqq\; \frac{\left(\frac{\rho}{c^2}\right)^{\rho/c^2}}{\Gamma\left(\frac{\rho}{c^2}\right)\bar{\gamma}\left(\frac{\rho}{c^2},\frac{\rho}{c^2}\right)} \frac{\Gamma\left(\frac{cs+v+\rho}{c^2}\right)\bar{\gamma}\left(\frac{cs+v+\rho}{c^2},\frac{v+\rho}{c^2}\right)}{\left(\frac{v+\rho}{c^2}\right)^{(cs+v+\rho)/c^2}} \exp\left\{\frac{v}{c^2}\right\}. \tag{72}$$

Then GP_α is a sub-Poisson uniform boundary with crossing probability α for scale c.

Proof of Proposition 10. The proof follows the same contours as that of Proposition 8. Using the sub-Poisson ψ function $\psi_P(\lambda) = c^{-2}(e^{c\lambda} - c\lambda - 1)$, the exponential integrand in our mixture is

$$\exp\left\{\lambda s - v\left(\frac{e^{c\lambda} - c\lambda - 1}{c^2}\right)\right\} = \theta^{(cs+v)/c^2} e^{(1-\theta)v/c^2}, \tag{73}$$

after substituting the one-to-one transformation $\theta = \theta(\lambda) \coloneqq e^{c\lambda}$, so that $\lambda = c^{-1}\log\theta$. We integrate against a gamma mixing distribution on θ with shape and scale parameters both equal to $\beta \coloneqq \rho/c^2$, truncated to $\theta \geq 1$, so that $\lambda \geq 0$:

$$e^{v/c^2}\frac{\int_1^\infty \theta^{(cs+v+\rho)/c^2-1}e^{-(v+\rho)\theta/c^2}\,\mathrm{d}\theta}{\int_1^\infty \theta^{\rho/c^2-1}e^{-\rho\theta/c^2}\,\mathrm{d}\theta} = \frac{\left(\frac{\rho}{c^2}\right)^{\rho/c^2}}{\Gamma\left(\frac{\rho}{c^2}\right)} \cdot \frac{\Gamma\left(\frac{cs+v+\rho}{c^2}\right)}{\left(\frac{v+\rho}{c^2}\right)^{(cs+v+\rho)/c^2}} \cdot \frac{\bar{\gamma}\left(\frac{cs+v+\rho}{c^2},\frac{v+\rho}{c^2}\right)}{\bar{\gamma}\left(\frac{\rho}{c^2},\frac{\rho}{c^2}\right)} \exp\left\{\frac{v}{c^2}\right\}. \tag{74}$$

This yields the closed-form mixture (72). To verify that our choice of β ensures that λ has approximate precision ρ under the full (not truncated) mixture distribution, we use the delta method to calculate the approximate variance of λ for large β based on the variance of θ under the full gamma mixture:

$$\mathrm{Var}\,\lambda \approx \left[\frac{1}{c^2\theta^2}\right]_{\theta=1} \cdot \frac{1}{\beta} = \frac{1}{\rho}. \tag{75}$$

□

A.4 Proof of Proposition 2

Under the conditions of Proposition 2, we have

$$m(s,v) = \int_0^{\lambda_{\max}} \exp\{\lambda s - \psi(\lambda)v\}\, f(\lambda)\,\mathrm{d}\lambda. \tag{76}$$

Note that $m(s,v)$ is nondecreasing in s and nonincreasing in v (since $\psi \geq 0$ by our assumptions on ψ).

Choose $\delta \in (0, \lambda_{\max})$ so that ψ has three continuous derivatives and f is continuous and positive on $[0,\delta)$; such a value of δ must exist by conditions (i) and (ii). Before proving Proposition 2, we state several lemmas.

Lemma 4. *Under the conditions of Proposition 2, for any $b \in (0, \psi'(\delta))$, we have $m(bv,v) < \infty$ and $m(bv,v) \to \infty$ as $v \to \infty$.*

Proof. Observe

$$m(bv, v) = \int_0^{\lambda_{\max}} \exp\{v[\lambda b - \psi(\lambda)]\} f(\lambda)\, \mathrm{d}\lambda. \tag{77}$$

Note $\frac{\mathrm{d}}{\mathrm{d}\lambda}[\lambda b - \psi(\lambda)] = b - \psi'(\lambda) < 0$ for all $\lambda \geq \delta$ by our condition on b. Hence the integrand $\exp\{v[\lambda b - \psi(\lambda)]\}$ is decreasing on $\lambda \geq \delta$ and bounded above by $e^{v\delta b}$ on $\lambda \leq \delta$ (since $\psi \geq 0$). The integrand is therefore uniformly bounded on $[0, \lambda_{\max})$, so that $m(bv, v) < \infty$.

Now Laplace's asymptotic approximation (Widder, 1942, Chapter VII.2, Theorem 2b) yields

$$\int_0^{\delta} \exp\{v\,[\lambda b - \psi(\lambda)]\}\, f(\lambda)\, \mathrm{d}\lambda \sim \frac{C e^{v\psi^\star(b)}}{\sqrt{v}}, \quad \text{as } v \to \infty, \tag{78}$$

where $C > 0$ is a constant not depending on v. (The condition $b < \psi'(\delta)$ ensures that the maximizer of $\lambda b - \psi(\lambda)$ lies within $[0, \delta)$.) Since the LHS of (78) lower bounds $m(bv, v)$ while the RHS diverges as $v \to \infty$, we must have $m(bv, v) \to \infty$ as $v \to \infty$. □

Lemma 5. *Under the conditions of Proposition 2, $m(\mathcal{M}_\alpha(v), v) = l_0/\alpha$ for all v sufficiently large.*

Proof. Let $\mathcal{C}(v) := [0, \psi'(\delta)v)$ for $v > 0$. Lemma 4 shows that $m(s, v) < \infty$ for all $s \in \mathcal{C}(v)$. Since $m(s, v)$ is nondecreasing in s, we may apply dominated convergence to find that $s \mapsto m(s, v)$ is continuous at all $s \in \mathcal{C}(v)$. Condition (i) of Proposition 2 implies $\psi \geq 0$, so that $m(0, v) \leq 1 \leq l_0/\alpha$ for all v. Finally, Lemma 4 shows that $\sup_{s \in \mathcal{C}(v)} m(s, v) \to \infty$ as $v \to \infty$. Hence, for v sufficiently large, there exists $s \in \mathcal{C}(v)$ such that $m(s, v) > l_0/\alpha$.

We have argued that, for any sufficiently large v, $m(0, v) \leq l_0/\alpha < m(\bar{s}, v) < \infty$ for some $\bar{s} < \psi'(\delta)v$, and $m(\cdot, v)$ is continuous on $[0, \bar{s}]$. The conclusion follows from the definition (13) of $\mathcal{M}_\alpha$. □

Lemma 6. *Under the conditions of Proposition 2, $\lim_{v\to\infty} \mathcal{M}_\alpha(v) = \infty$.*

Proof. Suppose for the sake of contradiction that there exists $a > 0$ such that $\mathcal{M}_\alpha(v) \leq a$ for all v. Then, since $m(s, v)$ is nondecreasing in s, $m(\mathcal{M}_\alpha(v), v) \leq m(a, v)$ for all v. But for sufficiently large v, we can write $a = bv$ for some $b < \psi'(\delta)$, so that Lemma 4 implies $m(a, v) < \infty$ for sufficiently large v. Since condition (i) of Proposition 2 implies $\psi \geq 0$, we have $m(s, v)$ is decreasing in v, and dominated convergence yields $m(a, v) \to 0$ as $v \to \infty$. But this implies $m(\mathcal{M}_\alpha(v), v) \to 0$, contradicting Lemma 5.

We have shown that $\limsup_{v\to\infty} \mathcal{M}_\alpha(v) = \infty$. But since $m(s, v)$ is nondecreasing in s and nonincreasing in v, Lemma 5 implies $\mathcal{M}_\alpha(v)$ must be nondecreasing in v. It follows that $\lim_{v\to\infty} \mathcal{M}_\alpha(v) = \infty$. □

Lemma 7. *Under the conditions of Proposition 2, $\mathcal{M}_\alpha(v) = o(v)$.*

Proof. Suppose for the sake of contradiction that $\mathcal{M}_\alpha(v) \geq bv$ for all v sufficiently large, for some $0 < b < \psi'(\delta)$. Then (again using the fact that $m(s, v)$ is nondecreasing in s) $\lim_{v\to\infty} m(\mathcal{M}_\alpha(v), v) \geq \lim_{v\to\infty} m(bv, v) = \infty$ by Lemma 4, contradicting Lemma 5. □

Proof of Proposition 2. We invoke Theorem 4 of Fulks (1951), setting Fulks' h equal to our v, Fulks' k equal to our $\mathcal{M}_\alpha(v)$, Fulks' ϕ equal to our ψ, Fulks' ψ equal to the identity function, Fulks' f equal to our f, and Fulks' b equal to our $\lambda_{\max}$. Fulks' assumptions (A1)-(A4) now read as follows.

(A1) requires $\psi(0) = \psi'(0_+) = 0$, $\psi''(0_+) > 0$, ψ has three continuous derivatives in a neighborhood of the origin, and ψ is positive and nondecreasing on $(0, \lambda_{\max})$.

(A2) requires conditions on the identity function which are trivially satisfied.

(A3) requires f to be integrable and to be continuous and positive at the origin.

(A4) requires $\mathcal{M}_\alpha(v) \to \infty$ as $v \to \infty$ and $\mathcal{M}_\alpha(v) = o(v)$.

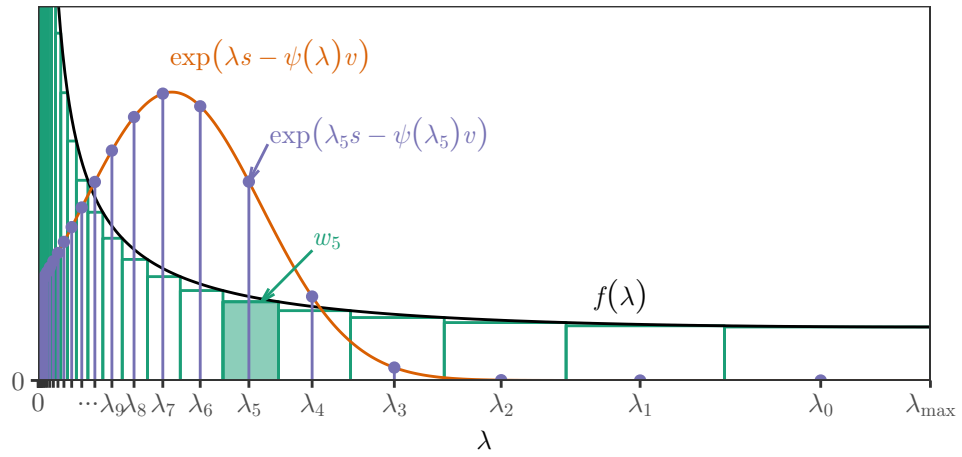

Figure 8: Illustration of Theorem 2. Mixture density $f(\lambda)$ is discretized on a grid $(\lambda_k)_{k=0}^{\infty}$ which gets finer as $\lambda \downarrow 0$. Resulting discrete mixture weights are represented by areas within green bars. Integrand $\exp\{\lambda s - \psi(\lambda)v\}$ is evaluated at grid points λ_k, illustrated by purple points. Multiplying one integrand evaluation $\exp\{\lambda_k s - \psi(\lambda_k)v\}$ by the corresponding weight w_k gives one term of the sum (17).

(A1) and (A3) are satisfied by conditions (i) and (ii) of Proposition 2. (A4) is satisfied by Lemmas 6 and 7. For Fulks' Theorem 4, it remains to verify that $\sqrt{v} = o(\mathcal{M}_\alpha(v))$. But if this were not true, then we could apply Theorem 1 or Theorem 2 of Fulks (1951) to conclude that $m(\mathcal{M}_\alpha(v), v) \to 0$ as $v \to \infty$, contradicting Lemma 5. So Fulks' Theorem 4 yields

$$m(\mathcal{M}_\alpha(v), v) \sim f(0)\sqrt{\frac{2\pi}{cv}} \exp\left\{\frac{\mathcal{M}_\alpha^2(v)}{2cv}\right\}. \tag{79}$$

Using Lemma 5 to set $m(\mathcal{M}_\alpha(v), v) = l_0/\alpha$, we may write

$$f(0)\sqrt{\frac{2\pi}{cv}} \exp\left\{\frac{\mathcal{M}_\alpha^2(v)}{2cv}\right\} = \frac{l_0 e^{o(1)}}{\alpha}, \tag{80}$$

which can be rearranged into the desired conclusion. □

We have proved the result for one-sided bounds, but a nearly-identical argument applies to two-sided bounds such as Proposition 7.

A.5 Proof of Theorem 2

Recall the discrete mixture support points and weights,

$$\lambda_k := \frac{\overline{\lambda}}{\eta^{k+1/2}} \quad \text{and} \quad w_k := \frac{\overline{\lambda}(\eta-1)f(\lambda_k\sqrt{\eta})}{\eta^{k+1}} \quad \text{for} \quad k = 0, 1, 2, \ldots. \tag{81}$$

Figure 8 illustrates the construction. To see heuristically why the exponentially-spaced grid $\lambda_k = \mathcal{O}(\eta^{-k})$ makes sense, observe that the integrand $\exp\{\lambda s - \lambda^2 v/2\}$ is a scaled normal density in λ with mean s/v and standard deviation $1/\sqrt{v}$. In the regime relevant to our curved boundaries, s is of order $\sqrt{v}$, ignoring logarithmic factors. Hence the integrand at time v has both center and spread of order $1/\sqrt{v}$, so as $v \to \infty$, the relevant scale of the integrand shrinks. With the grid $\lambda_k = \mathcal{O}(\eta^{-k})$ we have $\lambda_k - \lambda_{k+1} = \mathcal{O}(\lambda_k)$, ensuring that the resolution of the grid around the peak of the integrand matches the scale of the integrand as $v \to \infty$.

The discrete mixture bound is a valid mixture boundary in its own right, based on a discrete mixing distribution, but we may wish to know how well it approximates the continuous-mixture boundary from which it is derived. To illustrate the accuracy of the discrete mixture construction, we compare it to the one-sided normal mixture bound, Proposition 6. By using the same half-normal mixing density in Theorem 2 and setting $\eta = 1.05$, $\overline{\lambda} = 100$, we may evaluate a corresponding discrete mixture bound DM_α. With $\rho = 14.3$,

$\alpha = 0.05$ and $l_0 = 1$, numerical calculations indicate that $\mathrm{DM}_\alpha(v)/\mathrm{NM}_\alpha(v) \leq 1.004$ for $1 \leq v \leq 10^6$, suggesting that Theorem 2 gives an excellent conservative approximation to the corresponding continuous mixture boundary over a large practical range. Of course, when a closed form is available as in Proposition 6, one should use it in practice. But an exact closed form integral is rarely available as it is in Proposition 6, and substantial looseness often accompanies closed-form approximations which provably maintain crossing probability guarantees. In such cases, unless a closed form is required, Theorem 2 is preferable. See figure 10 for an example; in this figure, the bounds of Balsubramani (2014) and Darling and Robbins (1968) involve closed-form mixture integral approximations.

Proof of Theorem 2. Because f is nonincreasing, $f(\lambda) \geq f(\lambda_k\sqrt{\eta})$ on the interval $[\lambda_k/\sqrt{\eta}, \lambda_k\sqrt{\eta}]$, which has width $\lambda_{\max}(\eta-1)/\eta^{k+1} = w_k/f(\lambda_k\sqrt{\eta})$. Hence $\sum_{k=0}^{\infty} w_k \leq \int_0^\infty f(\lambda)\,\mathrm{d}\lambda = 1$. Let G be a discrete distribution which places mass $w_k/\sum_{j=0}^{\infty} w_j$ at the point λ_k. By Lemma 2, we know the mixture bound $\mathcal{M}_\alpha$ applied to the discrete mixture distribution G yields a sub-ψ uniform boundary with crossing probability α. But

$$\sum_{k=0}^{\infty} w_k \exp\left\{\lambda_k s - \psi(\lambda_k)v\right\} \leq \int \exp\left\{\lambda s - \psi(\lambda)v\right\} \mathrm{d}G(\lambda), \tag{82}$$

so $\mathrm{DM}_\alpha \geq \mathcal{M}_\alpha$. That is, our discrete mixture approximation DM_α is a conservative overestimate of a corresponding exact mixture boundary $\mathcal{M}_\alpha$, and can only have a lower crossing probability. So the discrete mixture bound DM_α satisfies the desired probability inequality $\mathbb{P}(\exists t : S_t \geq \mathrm{DM}_\alpha(V_t)) \leq \alpha$. □

A.6 Stitching as a discrete mixture approximation

Suppose we wish to analytically approximate the discrete mixture boundary DM_α of Theorem 2 in the sub-Gaussian case $\psi = \psi_N$. Clearly the sum is lower bounded by the maximum summand, which gives

$$\mathrm{DM}_\alpha(v) \leq \sup\left\{s \in \mathbb{R} : \sup_{k\geq 0}\left[w_k \exp\left\{\lambda_k s - \psi_N(\lambda_k)v\right\}\right] < \frac{l_0}{\alpha}\right\} \tag{83}$$

$$= \min_{k\geq 0}\left\{\frac{\log(l_0/w_k\alpha)}{\lambda_k} + \frac{\lambda_k}{2}\,v\right\}. \tag{84}$$

The last expression is the pointwise minimum of a collection of linear boundaries of the form presented in Lemma 1, each chosen with a different λ_k, and with nominal crossing rates $w_k\alpha$ so that a union bound over crossing events yields total crossing probability $\sum_k w_k\alpha \leq \alpha$. This is very similar to the stitching construction, with a slightly different choice of the sequence λ_k.

By equating w_k from Theorem 2 with $1/h(k)$ from Theorem 1, this observation allows us to view a stitched bound with function $h(k)$ as an approximation to a mixture bound with mixture density $f(\lambda) = \Theta(1/\lambda h(\log\lambda^{-1}))$ as $\lambda \downarrow 0$. For exponential stitching, this yields $f(\lambda) = \Theta(1)$—densities approaching a nonzero constant as $\lambda \downarrow 0$, including the half-normal distribution, correspond to exponential stitched boundaries growing at a rate $\sqrt{V_t \log V_t}$. For polynomial stitching, we have the corresponding mixture density

$$f_s^{\mathrm{LIL}}(\lambda) := \frac{(s-1)s^{s-1}1_{0\leq\lambda\leq\exp(-s)}}{\lambda\log^s\lambda^{-1}}, \tag{85}$$

matching the density from Balsubramani (2014, Lemma 12) (we truncate at $\lambda = e^{-s}$ to ensure the density is nonincreasing). The "slower" function $h(k) \propto k\log^s k$ corresponds to $f(\lambda) = \Theta(1/\lambda(\log\lambda^{-1})(\log\log\lambda^{-1})^s)$, the density from example 3 of Robbins (1970).

A.7 Proof of Theorem 3

The proof follows a straightforward idea. We break time into epochs $\eta^k \leq V_t < \eta^{k+1}$. Within each epoch we consider the linear boundary passing through the points $(\eta^k, g(\eta^k))$ and $(\eta^{k+1}, g(\eta^{k+1}))$. This line lies below $g(V_t)$ throughout the epoch, and its crossing probability is determined by its slope and intercept as in Lemma 1. Taking a union bound over epochs yields the result.

We need the following lemma concerning g:

Lemma 8. *If g is nonnegative and strictly concave on $\mathbb{R}_{\geq 0}$, then $g(v)$ is nondecreasing and $g(v)/v$ is strictly decreasing on $v > 0$.*

Proof. If $s < 0$ is a supergradient of g at some point t, then $g(t+u) < g(t) + su < 0$ for sufficiently large u, contradicting the non-negativity of g. So g is nondecreasing. Now fix $0 < x < y$ and let s be any supergradient of g at x. From nonnegativity and concavity we have $0 \leq g(0) \leq g(x) - xs$, so that $s \leq g(x)/x$. Strict concavity then implies $g(y) < g(x) + s(y-x) \leq g(x)y/x$. □

Proof of Theorem 3. Fix any $\eta > 1$. On $\eta^k \leq v < \eta^{k+1}$ we lower bound $g(v)$ by the line $a_k + b_k v$ passing through the points $(\eta^k, g(\eta^k))$ and $(\eta^{k+1}, g(\eta^{k+1}))$. This line has intercept and slope

$$a_k = \frac{\eta g(\eta^k) - g(\eta^{k+1})}{\eta - 1}, \tag{86}$$

$$b_k = \frac{g(\eta^{k+1}) - g(\eta^k)}{\eta^k(\eta - 1)}. \tag{87}$$

Note $a_k > 0$ and $b_k \geq 0$ by Lemma 8. We bound the upcrossing probability of this linear boundary using Lemma 1:

$$\mathbb{P}(\exists t \geq 1 : S_t \geq a_k + b_k V_t) \leq l_0 e^{-2a_k b_k} = l_0 \exp\left\{-\frac{2(g(\eta^{k+1}) - g(\eta^k))(\eta g(\eta^k) - g(\eta^{k+1}))}{\eta^k(\eta-1)^2}\right\}. \tag{88}$$

The conclusion follows from a union bound over epochs and from the arbitrary choice of η. □

Inspection of the proof reveals that the crossing probability bound (19) is valid not only for the boundary u given in (18), but also for a similar boundary which is finite and linear for all $v < 1$ and $v > v_{\max}$. This follows by extending the linear boundaries over the first and last epochs.

A.8 Proof of Theorem 4

For the proof, we take $a = 0, b = 1$ without loss of generality. Write $Y_t := X_t - \mathbb{E}_{t-1}X_t$ and $\delta_t := \widehat{X}_t - \mathbb{E}_{t-1}X_t$. Then $Y_t - \delta_t = X_t - \widehat{X}_t \in [-1, 1]$. We will show that $\exp\left\{\lambda \sum_{i=1}^t Y_i - \psi_E(\lambda)\sum_{i=1}^t (Y_i - \delta_i)^2\right\}$ is a supermartingale for each $\lambda \in [0, 1)$, where we take $c = 1$ in ψ_E.

The proof of Lemma 4.1 in Fan et al. (2015) shows that $\exp\left\{\lambda\xi - \psi_E(\lambda)\xi^2\right\} \leq 1 + \lambda\xi$ for all $\lambda \in [0, 1)$ and $\xi \geq -1$. Applied to $\xi = y - \delta$, we have

$$\exp\left\{\lambda y - \psi_E(\lambda)(y-\delta)^2\right\} \leq e^{\lambda\delta}(1 + \lambda(y - \delta)). \tag{89}$$

Since $Y_t - \delta_t \geq -1$, $\mathbb{E}_{t-1}Y_t = 0$, and δ_t is predictable, the above inequality implies

$$\mathbb{E}_{t-1}\exp\left\{\lambda Y_t - \psi_E(\lambda)(Y_t - \delta_t)^2\right\} \leq e^{\lambda\delta_t}(1 - \lambda\delta_t) \leq 1, \tag{90}$$

using $1 - x \leq e^{-x}$ in the final step.

This shows that $S_t = \sum_{i=1}^t Y_i = \sum_{i=1}^t X_i - t\mu_t$ is sub-exponential with variance process $V_t = \sum_{i=1}^t (Y_i - \delta_i)^2 = \sum_{i=1}^t (X_i - \widehat{X}_i)^2$ and scale $c = 1$. It follows that $\mathbb{P}(\exists t : S_t \geq u(V_t)) \leq \alpha$. A similar argument applied with $-X_t$ in place of X_t shows that $\mathbb{P}(\exists t : -S_t \geq u(V_t)) \leq \alpha$, and a union bound finishes the proof. □

We remark that the proofs of Maurer and Pontil (2009, Theorem 11), Audibert et al. (2009, Theorem 1), and Balsubramani and Ramdas (2016) follow very different arguments. All three proofs involve a Bennett-type concentration bound for the sample mean with a radius depending on the true variance, combined via a union bound with a concentration bound for the sample variance. Audibert et al. (2009) and Balsubramani and Ramdas (2016) achieve the latter bound using another Bennett/Bernstein-type inequality and the inequality $\mathbb{E}X^4 \leq \mathbb{E}X^2$ for $|X| \leq 1$, while Maurer and Pontil (2009) use a self-bounding property to achieve a concentration inequality for the sample variance directly (Maurer and Pontil, 2009, Theorem 7).

In contrast, our argument avoids the union bound over the sample mean and sample variance bounds. We achieve this by constructing an exponential supermartingale which directly relates the deviations of S_t to the "online" empirical variance V_t. In terms of proof technique, our method owes much more to the literature on self-normalized bounds (de la Peña, 1999, de la Peña et al., 2004, Bercu and Touati, 2008, Delyon, 2009 and especially Fan et al., 2015) than to the literature on empirical-Bernstein bounds.

A.9 Proof of Corollary 4

For case (1), Lemma 3(f) and Lemma 2 of Howard et al. (2020) (cf. Delyon, 2009) show that $S_t = \gamma_{\max}(Y_t)$ is sub-Gaussian with variance process $\widetilde{V}_t = \gamma_{\max}\left(\sum_{i=1}^t \frac{\Delta Y_i^2 + 2\mathbb{E}\Delta Y_i^2}{3}\right)$. Invoking Corollary 1, we have

$$\limsup_{t\to\infty} \frac{S_t}{\sqrt{2\widetilde{V}_t \log\log \widetilde{V}_t}} \le 1 \quad \text{a.s. on } \left\{\sup_t \widetilde{V}_t = \infty\right\}. \tag{91}$$

Applying the strong law of large numbers elementwise, we have $t^{-1}\sum_{i=1}^t \frac{\Delta Y_i^2 + 2\mathbb{E}\Delta Y_i^2}{3} \overset{\text{a.s.}}{\to} \mathbb{E}Y_1^2$ as $t \to \infty$, and the continuity of the maximum eigenvalue map over the set of positive semidefinite matrices ensures that $t^{-1}\widetilde{V}_t \overset{\text{a.s.}}{\to} \gamma_{\max}(\mathbb{E}Y_1^2) = t^{-1}V_t$. Hence, so long as $\mathbb{E}Y_1^2 > 0$ we conclude that, with probability one, $\sup_t \widetilde{V}_t = \infty$ and $\sqrt{\widetilde{V}_t \log\log \widetilde{V}_t} \sim \sqrt{\gamma_{\max}(\mathbb{E}Y_1^2) t \log\log t}$, completing the proof for case (1). (If $\mathbb{E}Y_1^2 = 0$ then the event $\{\sup_t V_t = \infty\}$ is empty and the result is vacuous.)

In case (2), Fact 1(d) and Lemma 2 of Howard et al. (2020) (cf. Tropp, 2012) show that (S_t) defined as above is sub-gamma with variance process (V_t) and scale c. The conclusion now follows directly from Corollary 1. □

A.10 Proof of Corollary 5

The argument is adapted from Tropp (2015). Let $X_i \coloneqq x_i x_i^T - \Sigma$. The triangle inequality implies $\|X_i\|_{\text{op}} \le \|x_i x_i^T\|_{\text{op}} + \|\Sigma\|_{\text{op}} \le 2b$. Hence, by Fact 1(c) and Lemma 2 of Howard et al. (2020) (cf. Tropp, 2012), $S_t = \gamma_{\max}\left(\sum_{i=1}^t X_i\right)$ is sub-Poisson with scale $c = 2b$ and variance process

$$\begin{aligned} V_t &= \gamma_{\max}\left(\sum_{i=1}^t \mathbb{E}X_i^2\right) && (92)\\ &= \gamma_{\max}\left(\sum_{i=1}^t \left[\mathbb{E}[(x_i x_i^T)^2] - \Sigma^2\right]\right) && (93)\\ &\le \sum_{i=1}^t \gamma_{\max}\left(\mathbb{E}[(x_i x_i^T)^2]\right). && (94)\end{aligned}$$

In the final step, we neglect the negative semidefinite term $-\Sigma^2$ and use the fact that the maximum eigenvalue of a sum of positive semidefinite matrices is bounded by the sum of the maximum eigenvalues. We continue by using $\|x_i x_i^T\| = \|x_i\|_2^2 \le b$ and the fact the expectation respects the semidefinite order to obtain

$$\begin{aligned} V_t &\le \sum_{i=1}^t \gamma_{\max}\left(\mathbb{E}\|x_i\|_2^2 x_i x_i^T\right) && (95)\\ &\le tb\|\Sigma\|_{\text{op}}. && (96)\end{aligned}$$

Plugging this upper bound on V_t into the discrete mixture bound of Theorem 2 gives the result. □

B Implications among sub-ψ boundaries

Together with Table 1, the following proposition formalizes the relationships illustrated in Figure 2, restating Proposition 2 of Howard et al. (2020) in the language of uniform boundaries. The first row of Table 1 uses the function

$$\varphi(g,h) \coloneqq \begin{cases} \frac{h^2-g^2}{2\log(h/g)}, & g < h \\ gh, & g \ge h. \end{cases} \tag{97}$$

	ψ_1	ψ_2	a	Restriction
(1)	ψ_N	$\psi_{B,g,h}$	$\frac{\varphi(g,h)}{gh}$	any $g, h > 0$
(2)	ψ_N	$\psi_{B,g,h}$	$\frac{(g+h)^2}{4gh}$	any $g, h > 0$
(3)	$\psi_{P,c}$	$\psi_{B,g,g+c}$	1	any $g > -c$
(5)	$\psi_{G,c}$	$\psi_{P,3c}$	1	
(6)	$\psi_{E,c}$	$\psi_{G,2c/3}$	1	
(7)	$\psi_{G,c}$	$\psi_{E,c}$	1	$c \geq 0$
(8)	$\psi_{G,c}$	$\psi_{E,2c}$	1	$c < 0$
(9)	$\psi_{P,c}$	$\psi_{G,c/2}$	1	$c < 0$
(10)	ψ_N	$\psi_{P,c}$	1	any $c < 0$
(11)	$\psi_{B,g,h}$	$\psi_{P,-g}$	1	

Table 1: For each row, if u is a sub-ψ_1 uniform boundary, subject to the given restriction, then $v \mapsto u(av)$ is a sub-ψ_2 uniform boundary. $\varphi(g,h)$ is defined in (97). See Proposition 11 for details.

Proposition 11. *For each row in Table 1, if u is a sub-ψ_1 uniform boundary, and the given restrictions are satisfied, then $v \mapsto u(av)$ is a sub-ψ_2 uniform boundary for the given constant a. Furthermore, when we allow only transformations of the form $v \mapsto u(av)$, these capture all possible implications among the five sub-ψ boundary types defined above, and the given constants are the best possible (in the case of row (2), the constant $(g+h)^2/4gh$ is the best possible of the form k/gh where k depends only on the total range $g+h$).*

A reader who is familiar with Howard et al. (2020) will note that the arrows in Figure 2 are reversed with respect to Figure 4 in their paper. Indeed, since any sub-Bernoulli process is also sub-Gaussian, it follows that any sub-Gaussian uniform boundary is also a sub-Bernoulli uniform boundary, and so on.

C Additional proofs

C.1 Proof of Proposition 3

Let $k \coloneqq (l_0/\alpha)^2$. For part (a), we will set the derivative of the squared objective $u^2(v)/v$ to zero:

$$\frac{\mathrm{d}}{\mathrm{d}v}\left[\left(1+\frac{\rho}{v}\right)\left(\log\left(\frac{k(v+\rho)}{\rho}\right)\right)\right] = -\frac{\rho}{v^2}\log\left(\frac{k(v+\rho}{\rho}\right) + \frac{1}{v} = 0. \tag{98}$$

$$-\left(\frac{v+\rho}{\rho}\right)\exp\left\{-\frac{v+\rho}{\rho}\right\} = -\frac{1}{ek}. \tag{99}$$

We solve this equation using the lower branch W_{-1} since we know $-(v+\rho)/\rho \leq -1$:

$$\frac{v+\rho}{\rho} = -W_{-1}\left(-\frac{1}{ek}\right), \tag{100}$$

which is equivalent to (21).

For part (b), we optimize the squared boundary $u^2(v)$:

$$\frac{\mathrm{d}}{\mathrm{d}\rho}\left[(v+\rho)\log\left(\frac{k(v+\rho)}{\rho}\right)\right] = \log\left(\frac{k(v+\rho)}{\rho}\right) - \frac{v}{\rho} = 0. \tag{101}$$

which is equivalent to (98). □

C.2 Proof of Proposition 4

First, Robbins and Siegmund (1970, Theorem 1) show that, for $B(t)$ a standard Brownian motion,

$$\mathbb{P}(\exists t \in (0,\infty) : B(t) \geq \mathcal{M}_\alpha(t)) = \alpha. \tag{102}$$

Let $(X_t)_{t=1}^\infty$ be any i.i.d. sequence of mean-zero random variables with unit variance and $\mathbb{E}e^{\lambda X_1} \le e^{\lambda^2/2}$, for example standard normal or Rademacher random variables. For each $m \in \mathbb{N}$, let $S_t^{(m)} := \sum_{i=1}^t X_i/\sqrt{m}$ and $V_t^{(m)} := t/m$, noting that $(S_t^{(m)})$ is sub-Gaussian with variance process $(V_t^{(m)})$. Our proof rests upon a standard application of Donsker's theorem, detailed below, which shows that, for any $T \in \mathbb{N}$,

$$\lim_{m\to\infty} \mathbb{P}\left(\exists t \in [mT] : S_t^{(m)} \ge \mathcal{M}_\alpha(V_t^{(m)})\right) = \mathbb{P}(\exists t \in (0,T] : B(t) \ge \mathcal{M}_\alpha(t)). \tag{103}$$

To obtain the desired conclusion from (103), we write, for any $m \in \mathbb{N}$ and $T \in \mathbb{N}$,

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t^{(m)} \ge \mathcal{M}_\alpha(V_t^{(m)})\right) \ge \mathbb{P}\left(\exists t \in [mT] : S_t^{(m)} \ge \mathcal{M}_\alpha(V_t^{(m)})\right). \tag{104}$$

Take $m \to \infty$ and use (103) to find, for any $T \in \mathbb{N}$,

$$\liminf_{m\to\infty} \mathbb{P}\left(\exists t \in \mathbb{N} : S_t^{(m)} \ge \mathcal{M}_\alpha(V_t^{(m)})\right) \ge \mathbb{P}\left(\exists t \in (0,T] : B(t) \ge \mathcal{M}_\alpha(t)\right). \tag{105}$$

Now take $T \to \infty$ to obtain

$$\liminf_{m\to\infty} \mathbb{P}\left(\exists t \in \mathbb{N} : S_t^{(m)} \ge \mathcal{M}_\alpha(V_t^{(m)})\right) \ge \mathbb{P}\left(\exists t \in (0,\infty) : B(t) \ge \mathcal{M}_\alpha(t)\right) = \alpha, \tag{106}$$

by (102). But for each $m \in \mathbb{N}$, $S_t^{(m)}$ is sub-Gaussian with variance process $V_t^{(m)}$, so that

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t^{(m)} \ge \mathcal{M}_\alpha(V_t^{(m)})\right) \le \alpha. \tag{107}$$

Together, (106) and (107) yield

$$\lim_{m\to\infty} \mathbb{P}\left(\exists t \in \mathbb{N} : S_t^{(m)} \ge \mathcal{M}_\alpha(V_t^{(m)})\right) = \alpha. \tag{108}$$

Since $(S_t^{(m)}, V_t^{(m)}) \in \mathbb{S}^1_{\psi_N}$ for each m, the conclusion follows.

To prove (103), we will use the fact that $\mathcal{M}_\alpha : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ is continuous, increasing and concave, as proved in Lemma 9 below. For each $t \in \mathbb{R}_{>0}$ let $S(mt)$ be equal to S_{mt} for $mt \in \mathbb{N}$ and a linear interpolation otherwise (with $S(0) = 0$). Let $C[0,T]$ denote the space of continuous, real-valued functions on $[0,T]$ equipped with the sup-norm, and let $\mathbb{P}_0$ denote the probability measure for standard Brownian motion. We first use a corollary of Donsker's theorem: for any $\varphi : C[0,T] \to \mathbb{R}$ continuous $\mathbb{P}_0$-a.s., we have (Durrett, 2017, Theorems 8.1.5, 8.1.11)

$$\varphi\left(\frac{S(m\cdot)}{\sqrt{m}}\right) \xrightarrow{\mathrm{d}} \varphi(B(\cdot)) \quad \text{as } m \to \infty. \tag{109}$$

We let $\varphi(f) := \sup_{t\in[0,T]}[f(t) - \mathcal{M}_\alpha(t)]$, so that by compactness of $[0,T]$ and continuity of f and $\mathcal{M}_\alpha$, $\varphi(f) \ge 0$ if and only if $f(t) \ge \mathcal{M}_\alpha(t)$ for some $t \in [0,T]$. Now $\varphi(S(m\cdot)/\sqrt{m}) \xrightarrow{\mathrm{d}} \varphi(B(\cdot))$, and note that $\varphi(B(\cdot))$ has a continuous distribution: the distribution when $\mathcal{M}_\alpha(t) \equiv 0$ is well-known by the reflection principle, and the measure for the Brownian motion with drift $B(t) - \mathcal{M}_\alpha(t) + \mathcal{M}_\alpha(0)$ is equivalent to the measure for $B(t)$ by the Cameron-Martin theorem (Morters and Peres, 2010, Theorem 1.38). Hence

$$\mathbb{P}\left(\exists t \in [0,T] : \frac{S(mt)}{\sqrt{m}} \ge \mathcal{M}_\alpha(t)\right) \to \mathbb{P}\left(\exists t \in [0,T] : B(t) \ge \mathcal{M}_\alpha(t)\right). \tag{110}$$

But because $\mathcal{M}_\alpha(t)$ is concave, the linear interpolation of $S(\cdot)$ cannot add any new upcrossings beyond those in (S_t):

$$\mathbb{P}\left(\exists t \in [0,T] : \frac{S(mt)}{\sqrt{m}} \ge \mathcal{M}_\alpha(t)\right) = \mathbb{P}\left(\exists x \in [mT] : \frac{S_x}{\sqrt{m}} \ge \mathcal{M}_\alpha(x/m)\right) \tag{111}$$

$$= \mathbb{P}\left(\exists t \in [mT] : S_t^{(m)} \ge \mathcal{M}_\alpha(V_t^{(m)})\right). \tag{112}$$

Combining (112) with (110) yields (103), completing the proof. □

Lemma 9. *The function $\mathcal{M}_\alpha : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ is continuous, increasing and concave.*

Proof. Continuity of $\mathcal{M}_\alpha(v)$ is clear from the continuity of $\exp\{\lambda s - \psi(\lambda)v\}$ in s and v, which also implies

$$\int \exp\{\lambda \mathcal{M}_\alpha(v) - \psi(\lambda)v\}\, \mathrm{d}F(\lambda) = \frac{l_0}{\alpha} \tag{113}$$

for all $v > 0$. That is, the left-hand side is constant in v, hence has derivative with respect to v equal to zero. We may exchange the derivative and integral by Theorem A.5.1 of Durrett (2017), noting that the integrand is positive and continuously differentiable in v and F is a probability measure. This yields

$$\mathcal{M}'_\alpha(v) = \frac{A(v)}{B(v)} > 0, \tag{114}$$

$$\text{where } A(v) := \int \psi(\lambda) e^{\lambda \mathcal{M}_\alpha(v) - \psi(\lambda)v}\, \mathrm{d}F(\lambda) \tag{115}$$

$$\text{and } B(v) := \int \lambda e^{\lambda \mathcal{M}_\alpha(v) - \psi(\lambda)v}\, \mathrm{d}F(\lambda). \tag{116}$$

Both $A(v) > 0$ and $B(v) > 0$ since the integrands are positive, which shows that $\mathcal{M}_\alpha$ is increasing. Differentiating again yields, after some algebra,

$$B^2(v)\mathcal{M}''_\alpha(v) = \int \left(-\frac{[\lambda A(v) - \psi(\lambda)B(v)]^2}{B(v)} \right) e^{\lambda \mathcal{M}_\alpha(v) - \psi(\lambda)v}\, \mathrm{d}F(\lambda) \le 0, \tag{117}$$

since the integrand is now nonpositive, showing that $\mathcal{M}_\alpha$ is concave. □

C.3 Proof of Corollary 6

Write $\mu^\star := \mathbb{E}T(X_1)$.We have noted in the discussion preceding the result that the exponential process $\exp\{\lambda S_t(\mu) - t\psi_\mu(\lambda)\}$ is the likelihood ratio testing $H_0 : \theta = \theta(\mu)$ against $H_1 : \theta = \theta(\mu) + \lambda$. It is well-known that the likelihood ratio is a martingale under the null. Hence $(S_t(\mu^\star))$ is sub-$\psi_{\mu^\star}$ with variance process $V_t = t$, and it follows immediately that $\mathbb{P}(\exists t : S_t(\mu^\star) \ge u_{\mu^\star}(t)) \le \alpha_1$. Apply the same argument with $-X_t$ in place of X_t to conclude that $\mathbb{P}(\exists t : -S_t(\mu^\star) \ge \tilde{u}_{\mu^\star}(t)) \le \alpha_2$. A union bound completes the argument. □

C.4 Proof of Lemma 3

The implication $(a) \Rightarrow (b)$ follows from

$$A_T = \left[\bigcup_{t=1}^{\infty} A_t \cap \{T = t\} \right] \cup [A_\infty \cap \{T = \infty\}] \subseteq \bigcup_{t=1}^{\infty} A_t. \tag{118}$$

It is clear that $(b) \Rightarrow (c)$. For $(c) \Rightarrow (a)$, take $\tau = \inf\{t \in \mathbb{N} : A_t \text{ occurs}\}$, so that $A_\tau = \bigcup_{t=1}^{\infty} A_t$. □

D Computing conjugate mixture bounds by root-finding

In this section we demonstrate that our conjugate mixture boundaries, which involve the supremum $\mathcal{M}_\alpha(v)$ defined in (13), can be computed via root-finding. We assume that ψ is CGF-like, a property which holds for all of the ψ functions in Section 2:

Definition 3 (Howard et al., 2020, Definition 2)**.** A real-valued function ψ with domain $[0, \lambda_{\max})$ is called *CGF-like* if it is strictly convex and twice continuously differentiable with $\psi(0) = \psi'(0_+) = 0$ and $\sup_{\lambda \in [0,\lambda_{\max})} \psi(\lambda) = \infty$. For such a function, we write

$$\bar{b} := \sup_{\lambda \in [0,\lambda_{\max})} \psi'(\lambda) \in (0, \infty]. \tag{119}$$

Lemma 2 implies that, with probability at least $1-\alpha$, $m(S_t, V_t) < l_0/\alpha$ for all t, where

$$m(s,v) = \int \exp\{\lambda s - \psi(\lambda)v\}\, \mathrm{d}F(\lambda). \tag{120}$$

We are interested in the set $A(v) \coloneqq \{s \in \mathbb{R} : m(s,v) < l_0/\alpha\}$ for fixed $v \geq 0$. It is clear that $m(0,v) \leq 1 < l_0/\alpha$ whenever $l_0 \geq 1$ (which holds in all cases we consider), since $\psi \geq 0$, $v \geq 0$ and F is a probability distribution. So $0 \in A(v)$ always. We show below that, in addition, $A(v)$ is always an interval.

For one-sided boundaries, F is supported on $\lambda \geq 0$, and so long as F is not a point mass at zero (which would be an uninteresting mixture), $m(s,v)$ is strictly increasing in s whenever $m(s,v) < \infty$. Hence $m(s,v) = l_0/\alpha$ for at most one value of $s^\star(v) > 0$, in which case $A(v) = (-\infty, s^\star(v))$.

It is possible that $m(s,v) < l_0/\alpha$ for all s where the integral converges. To examine this case, we fix $v > 0$, which is the interesting case in practice, and make two observations:

- Whenever $s < \bar{b}v$, we have $m(s,v) < \infty$. Indeed, in this case, $\exp\{\lambda s - \psi(\lambda)v\} \to 0$ as $\lambda \to \infty$, and as the integrand is continuous in λ, it must be uniformly bounded. It follows immediately that we can have $m(s,v) = \infty$ only when $\bar{b} < \infty$.
- Whenever $\bar{b} < \infty$, we have $S_t \leq \bar{b}V_t$ a.s., a consequence of Theorem 1(a) of Howard et al. (2020), which shows that $\mathbb{P}(\exists t : S_t \geq a + \bar{b}V_t) = 0$ for all $a > 0$. (To verify this fact, note we must have $\lambda_{\max} = \infty$ when $\bar{b} < \infty$ in order for the CGF-like condition $\sup_{\lambda \in [0, \lambda_{\max})} \psi(\lambda) = \infty$ to hold.)

Hence, when $\bar{b} = \infty$ we need not worry about $m(s,v) = \infty$. When $\bar{b} < \infty$, it suffices to check $m(\bar{b}v, v)$, which may be infinite. If $m(\bar{b}v, v) \geq l_0/\alpha$, then we search for a root of $m(s,v) = l_0/\alpha$ in the interval $s \in [0, \bar{b}v]$. If $m(\bar{b}v, v) < l_0/\alpha$, it suffices to take $\mathcal{M}_\alpha(v) = \bar{b}v + \epsilon$ for any $\epsilon > 0$. In practice, it seems more reasonable to take the upper bound $\bar{b}v$ and use a closed confidence set instead of an open one.

For two-sided boundaries, when F has support on both $\lambda > 0$ and $\lambda < 0$, in general we require the technical condition

$$\int |\lambda|^k \exp\{\lambda s - \psi(\lambda)v\}\, \mathrm{d}F(\lambda) < \infty, \quad \text{for } k = 1, 2. \tag{121}$$

This ensures that we may differentiate $m(s,v)$ twice with respect to s, exchanging the derivative and the integral both times (Durrett, 2017, Theorem A.5.3). Hence, whenever condition (121) holds,

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2} m(s,v) = \int \lambda^2 \exp\{\lambda s - \psi(\lambda)v\}\, \mathrm{d}F(\lambda) \geq 0, \tag{122}$$

so that $m(s,v)$ is convex in s for each $v \geq 0$. As $m(0,v) < l_0/\alpha$, we conclude that $m(s,v) = l_0/\alpha$ for at most one value $s^\star(v) > 0$ and one value $s_\star(v) < 0$, and $A(v) = (s_\star(v), s^\star(v))$. A similar discussion as above applies when $\bar{b} < \infty$ and we may have $m(s,v) = \infty$ for some values of s.

As Proposition 5 yields a closed-form result, only Proposition 7 requires that we verify condition (121). From the proof of Proposition 7 in Appendix A.3, it suffices to show that

$$\int_0^1 \left| \log\left(\frac{p}{1-p}\right) \right|^k p^a (1-p)^b \, \mathrm{d}p < \infty \tag{123}$$

for some $a, b > 0$ and $k = 1, 2$. This follows from the fact that the integrand is continuous on $p \in (0,1)$ and approaches zero as $p \to 0$ and $p \to 1$, so it is bounded.

E Tuning discrete mixture implementation

In Section 3.5 we have discussed the choice of mixing precision in order to tune a mixture bound for a particular range of sample sizes. For discrete mixtures, the value $\overline{\lambda}$ must also be chosen, and this depends on the minimum relevant value of V_t: making $\overline{\lambda}$ larger will make the resulting bound tighter over smaller values of V_t at the cost of a looser bound for larger values of V_t. In practice, for $\psi = \psi_G$, setting $\overline{\lambda} =$

$[c + \sqrt{m/2 \log \alpha^{-1}}]^{-1}$ will ensure the bound is tight for $V_t \geq m$. Furthermore, when evaluating $\mathrm{DM}_\alpha(v)$ in practice, the sum can be truncated after $k_{\max} = \lceil \log_\eta(\overline{\lambda}[c + \sqrt{5v/\log \alpha^{-1}}]) \rceil$ terms. The remainder of this section explains these choices.

We wish to understand what range of values of λ our discrete mixture must cover to ensure we get a tight bound for all $V_t \in [m, v_{\max}]$. At $V_t = m$ the value of λ which yields the optimal linear bound from Lemma 1 is found by optimizing

$$\frac{\log \alpha^{-1}}{\lambda} + \frac{\psi(\lambda)}{\lambda} \cdot m, \tag{124}$$

yielding the first-order condition

$$\lambda \psi'(\lambda) - \psi(\lambda) = \frac{\log \alpha^{-1}}{m}. \tag{125}$$

For $\psi = \psi_G$, this becomes

$$\frac{\lambda^2}{2(1 - c\lambda)^2} = \frac{\log \alpha^{-1}}{m}, \tag{126}$$

which is solved by

$$\lambda^\star(m) = \frac{1}{c + \sqrt{m/2 \log \alpha^{-1}}}. \tag{127}$$

Large values of λ are necessary to achieve tight bounds for small V_t. Hence, to ensure good performance at $V_t = m$ we choose $\overline{\lambda} = [c + \sqrt{m/2 \log \alpha^{-1}}]^{-1}$. Similarly, to ensure the sum safely covers $V_t = v$ we ensure $\lambda_{k_{\max}} \leq [c + \sqrt{10v/2 \log \alpha^{-1}}]^{-1}$ (using an arbitrary "fudge factor" of ten), which yields $k_{\max} = \lceil \log_\eta(\lambda_{\max}[c + \sqrt{5v/\log \alpha^{-1}}]) \rceil$.

We note that η must also be chosen, but the only tradeoff here is computational. Smaller values of η lead to more accurate approximations of the discrete mixture to the target continuous mixture, but require more terms to be summed. We have found $\eta = 1.1$ to provide excellent approximations in the examples we have examined.

F Intrinsic time, change of units and minimum time conditions

In this section we point out that a bound expressed in terms of intrinsic time yields an infinite family of related bounds via scaling, and that "minimum time" conditions in such bounds (such as $m \vee V_t$ in Theorem 1) can be freely scaled as well. Suppose we have a uniform bound of the form

$$\mathbb{P}\left(\exists t \geq 1 : S_t \geq u_c(m \vee V_t)\right) \leq \alpha, \tag{128}$$

where intrinsic time V_t has the same units as S_t^2, as usual, and c is some parameter with the same units as S_t. Then, fixing any $\gamma > 0$ and applying the bound (128) to the scaled observations $X_t/\sqrt{\gamma}$, which amounts to a change of units, we have

$$\alpha \geq \mathbb{P}\left(\exists t \geq 1 : \frac{S_t}{\sqrt{\gamma}} \geq u_{c/\sqrt{\gamma}}\left(m \vee \frac{V_t}{\gamma}\right)\right) \tag{129}$$

$$= \mathbb{P}\left(\exists t \geq 1 : S_t \geq h_c(\gamma m \vee V_t)\right), \quad \text{where } h_c(v) \coloneqq \sqrt{\gamma} u_{c/\sqrt{\gamma}}\left(\frac{v}{\gamma}\right). \tag{130}$$

By changing units we have obtained a new bound on S_t with different minimum time γm and a different shape. For example, applying this change of units to the stitched boundary (8) with $m = 1$ yields the family of bounds

$$\mathbb{P}\left(\exists t \geq 1 : S_t \geq k_1 \sqrt{(\gamma \vee V_t)\ell\left(\frac{\gamma \vee V_t}{\gamma}\right)} + c k_2 \ell\left(\frac{\gamma \vee V_t}{\gamma}\right)\right) \leq \alpha \tag{131}$$

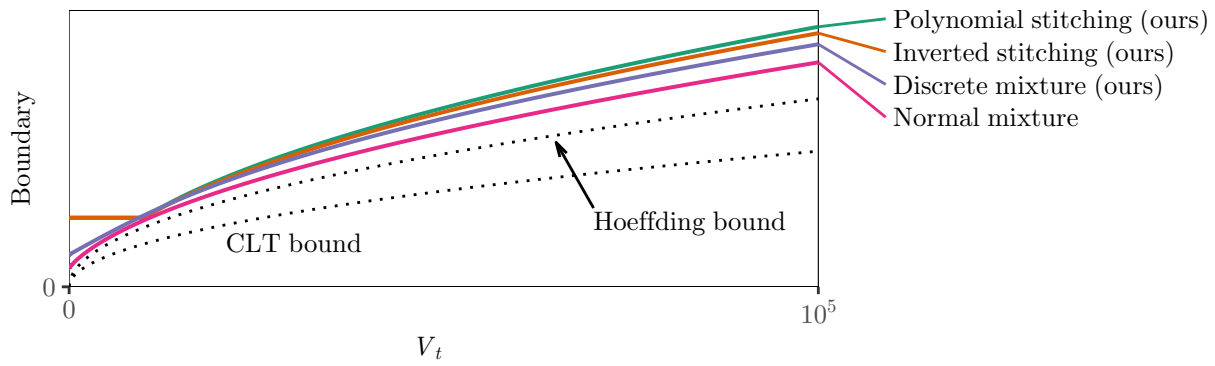

Figure 9: Pointwise and uniform bounds for independent 1-sub-Gaussian observations, $\alpha = 0.025$. All tuning parameters are chosen to optimize roughly for time $V_t = 10^4$. The dotted lines show the Hoeffding bound $\sqrt{2V_t \log \alpha^{-1}}$, which is nonasymptotically pointwise valid, and the CLT bound $z_{1-\alpha}\sqrt{V_t}$, which is asymptotically pointwise valid. Polynomial stitching uses Theorem 1 with $\eta = 2.04$, $m = 10^4$, and $h(k) = (k+1)^{1.4}\zeta(1.4)$. The inverted stitching boundary is $1.7\sqrt{(V_t \vee 10^4)(\log(1 + \log((V_t \vee 10^4)/10^4) + 3.5)}$, using Theorem 3 with $\eta = 2.99$, $v_{\max} = 10^{20}$, and error rate 0.82α to account for finite horizon and ensure a fair comparison. Discrete mixture applies Theorem 2 to the density $f(\lambda) = 0.4 \cdot 1_{0 \le \lambda \le \lambda_{\max}} / [\lambda \log^{1.4}(\lambda_{\max} e/\lambda)]$ with $\eta = 1.1$ and $\lambda_{\max} = 0.044$; see Appendix A.6 for motivation. The normal mixture bound (53) uses $\rho = 1260$.

for any $\gamma > 0$, with the definition of ℓ unchanged from (8). Note only the argument of ℓ has been scaled. We started with a single bound (8) expressed in terms of V_t and ended up with a family of bounds on the same process S_t, one for each value of γ. Indeed, the tuning parameter m in Theorem 1 is obtained by exactly this argument. The effect is more clear if we let $c = 0$ and examine the upper bound on the normalized process $S_t/\sqrt{V_t}$: then for any $\gamma > 0$, with probability at least $1 - \alpha$,

$$\frac{S_t}{\sqrt{V_t}} \le \begin{cases} k_1 \sqrt{\ell\left(\frac{V_t}{\gamma}\right)}, & \text{when } V_t \ge \gamma, \\ k_1 \sqrt{\frac{\gamma \ell(1)}{V_t}}, & \text{when } V_t < \gamma. \end{cases} \tag{132}$$

Now the right-hand depends on V_t only through V_t/γ, so that the effect of changing γ is simply to multiplicatively shift the bound backwards or forwards in time without changing the bounded process.

G Detailed comparison of finite LIL bounds

Figures 9 and 10 compare our finite LIL bounds to several existing bounds. Below we restate the original results from the various papers giving finite LIL bounds included in Figure 10. In table 2, for ease of comparison, we write all bounds in the form

$$\mathbb{P}(\exists t \ge 1 : S_t \ge A\sqrt{t(\log\log Bt + C)}, \tag{133}$$

valid for independent 1-sub-Gaussian observations. When the original bound holds only for $t \ge n$ instead of $t \ge 1$, we apply a change of units argument to replace $\log\log Bt$ with $\log\log Bnt$ and $t \ge n$ with $t \ge 1$, so that all bounds are comparable (see Appendix F). When bounds are expressed in terms of intrinsic time V_t (Balsubramani, 2014), this is formally justified. When they are expressed in terms of nominal time (Darling and Robbins, 1967*b*, 1968) this is only a heuristic argument, but we conjecture that proofs of such bounds could be generalized to justify this scaling. When observations are i.i.d. from an infinitely divisible distribution, the change is formally justified by replacing each observation X_i with a sum of n i.i.d. "pseudo-observations" Z_i such that $\sum_{i=1}^n Z_i \sim X_1$.

- Jamieson and Nowak (2014), Lemma 1: for i.i.d. sub-Gaussian observations with variance parameter σ^2,

$$\mathbb{P}\left(\exists t \ge 1 : S_t \ge (1+\sqrt{\epsilon})\sqrt{2\sigma^2(1+\epsilon)t\log\left(\frac{\log((1+\epsilon)t)}{\delta}\right)}\right) \le 1 - \frac{2+\epsilon}{\epsilon}\left(\frac{\delta}{\log(1+\epsilon)}\right)^{1+\epsilon}. \tag{134}$$

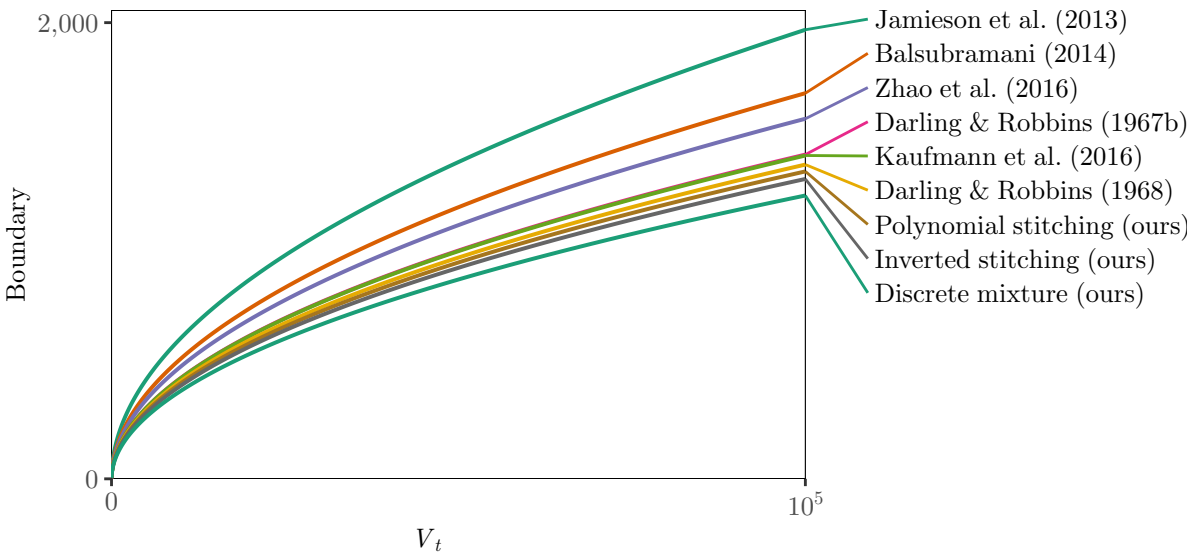

Figure 10: Finite LIL bounds for independent 1-sub-Gaussian observations, $\alpha = 0.025$. The dotted lines show the Hoeffding bound $\sqrt{2V_t \log \alpha^{-1}}$, which is nonasymptotically pointwise valid, and the CLT bound $z_{1-\alpha}\sqrt{V_t}$, which is asymptotically pointwise valid. Polynomial stitching uses Theorem 1 with $\eta = 2.04$ and $h(k) = (k+1)^{1.4}\zeta(1.4)$. The inverted stitching boundary is $1.7\sqrt{V_t(\log(1+\log V_t)+3.5)}$, using Theorem 3 with $\eta = 2.99$, $v_{\max} = 10^{20}$, and error rate 0.82α to account for finite horizon. Discrete mixture applies Theorem 2 to the density $f(\lambda) = 0.4 \cdot 1_{0\le\lambda\le 4}/[\lambda \log^{1.4}(4e/\lambda)]$ with $\eta = 1.1$, and $\lambda_{\max} = 4$; see Appendix A.6 for motivation. The normal mixture bound (53) uses $\rho = 0.129$. See Appendix G for details.

- Zhao et al. (2016), Theorem 1: for sub-Gaussian observations with variance parameter $1/4$,

$$\mathbb{P}\left(\exists t \ge 1 : S_t \ge \sqrt{at \log(\log_c t + 1) + bt}\right) \le \zeta(2a/c)e^{-2b/c}. \tag{135}$$

- Kaufmann et al. (2016), Lemma 7: for independent sub-Gaussian observations with variance parameter σ^2,

$$\mathbb{P}\left(\exists t \ge 1 : S_t \ge \sqrt{2\sigma^2 t(x + \eta \log\log(et))}\right) \le \sqrt{e}\zeta\left(\eta\left(1 - \frac{1}{2x}\right)\right)\left(\frac{\sqrt{x}}{2\sqrt{2}} + 1\right)^{\eta} e^{-x} \tag{136}$$

- Balsubramani (2014), Theorem 4: for $|X_t| \le c_t$ a.s. and $V_t = \sum_{i=1}^t c_i^2$,

$$\mathbb{P}\left(\exists t \ge 1 : V_t \ge 173 \log\left(\frac{2}{\alpha}\right) : S_t \ge \sqrt{3V_t(2\log\log(3V_t/2S_t) + \log \alpha^{-1})}\right) \le \alpha. \tag{137}$$

  Though the bound is stated for bounded observations, the proof holds for any observations sub-Gaussian with variance parameters (c_t^2), as noted in section 5.2 of Balsubramani (2014). Balsubramani suggests removing the initial time condition by imposing a constant bound over $t \le 173\log(2/\alpha)$ (section 5.3). We instead remove the condition by a change of units, as discussed in Appendix F.

- Darling and Robbins (1967*b*), eq. 22: for i.i.d. observations sub-Gaussian with variance parameter 1,

$$\mathbb{P}\left(\exists t \ge \eta^j : S_t \ge \frac{1+\eta}{2\sqrt{\eta}}\sqrt{t(2c\log\log t - 2c\log\log\eta + 2\log a)}\right) \le \frac{1}{a(c-1)(j-1/2)^{c-1}}. \tag{138}$$

  Darling and Robbins consider results for a general bound $\varphi(\lambda)$ on the moment-generating function of the observations. The result involves the term $h(v_t)$ where the function $h(\lambda) := 1/2 + \lambda^{-2}\log\varphi(\lambda)$ and v_t is unspecified but bounded.

- Darling and Robbins (1968), eq. 2.2 and the example that follows: for i.i.d. observations sub-Gaussian with variance parameter 1,

$$\mathbb{P}\left(\exists t \ge 3 : S_t \ge A\sqrt{t(\log\log t + C)}\right) \le \int_m^\infty \frac{A\sqrt{\log\log t + C}}{t} \exp\left\{-\frac{A^2(\log\log t + C)}{2}\right\} \mathrm{d}t. \tag{139}$$

Darling and Robbins give a closed-form upper bound for the right-hand side of (139). We instead evaluate it numerically, using readily-available implementations of the upper incomplete gamma function:

$$\int_m^\infty \frac{A\sqrt{\log\log t + C}}{t} \exp\left\{-\frac{A^2(\log\log t + C)}{2}\right\} \mathrm{d}t = \frac{\sqrt{2\pi}Ae^{-C}}{(A-2)^{3/2}} \mathbb{P}\left(G \geq \frac{A^2-2}{2}(\log\log m + C)\right), \quad (140)$$

where $G \sim \Gamma(3/2, 1)$.

- Polynomial stitching as in (10) with $c = 0$.
- Inverted stitching with $g(v) = A\sqrt{v(\log\log(ev) + C)}$ as in (20). We set $v_{\max} = 10^{20}$ which covers 42 epochs with $\eta = 2.994$. To make for a fair comparison with polynomial stitching, observe that in 42 epochs with $s = 1.4$, polynomial stitching "spends" $\sum_{k=1}^{42} k^{-1.4}/\zeta(1.4) \approx 0.820$ of its crossing probability α, so we run inverted stitching with $\alpha = 0.820 \cdot 0.025$.
- Normal mixture as in (53) with $\rho \approx 0.13$:

$$u(v) \approx \sqrt{2(v+0.13)\log\left(20\sqrt{1+\frac{v}{0.13}}+1\right)}. \quad (141)$$

This is not a LIL boundary, so is not included in Table 2.

Source and parameter settings	A	B	C
Jamieson and Nowak (2014)	$(1+\sqrt{\epsilon})\sqrt{2(1+\epsilon)}$	$1+\epsilon$	$\frac{1}{1+\epsilon}\log\left(\frac{2+\epsilon}{\alpha\epsilon\log^{1+\epsilon}(1+\epsilon)}\right)$
$\epsilon = 0.033$	(1.7)	(1.033)	(10.966)
Balsubramani (2014)	$\sqrt{6}$	$\frac{865}{2}\log\left(\frac{2}{\delta}\right)$	$(\log\alpha^{-1})/2$
	(2.45)	(1137)	(1.844)
Zhao et al. (2016)	$2\sqrt{a}$	c	$\frac{c}{2a}\log\left(\frac{\zeta(2a/c)}{\alpha\log^{2a/c}c}\right)$
$a = 0.7225, c = 1.1$	(1.7)	(1.1)	(6.173)
Darling and Robbins (1967*b*)	$(1+\eta)\sqrt{\frac{c}{2\eta}}$	η^j	$\frac{1}{c}\log\left(\frac{1}{\alpha(c-1)(j-1/2)^{c-1}\log^c\eta}\right)$
$j = 1, c = 1.4, \eta = 1.429$	(1.7)	(1.429)	(4.518)
Kaufmann et al. (2016)	$\sqrt{2\eta}$	e	$x(\alpha,\eta)/\eta$
$\eta = 1.3$	(1.7)	(2.718)	(4.427)
Darling and Robbins (1968)	A	3	$C(\alpha, A)$
$A = 1.7$	(1.7)	(3)	(3.945)
Polynomial stitching (10)	$(\eta^{1/4}+\eta^{-1/4})\sqrt{\frac{s}{2}}$	η	$\frac{1}{s}\log\frac{\zeta(s)}{\alpha\log^s\eta}$
$s = 1.4, \eta = 2.041$	(1.7)	(2.041)	(3.782)
Inverted stitching (Theorem 3)	A	e	$C(\alpha, A, \eta)$
$\eta = 2.994$, nominal error rate 0.82α	(1.7)	(2.718)	(3.454)

Table 2: Comparison of parameters A, B, C for finite LIL boundaries expressed in the form $\mathbb{P}(\exists t \geq 1 : S_t \geq A\sqrt{t(\log\log Bt + C)}) \leq \alpha$ for sums of independent 1-sub-Gaussian observations, with $\alpha = 0.025$. Functions $x(\alpha, \eta)$ and $C(\alpha, \dots)$ are given by numerical root-finding to set the corresponding error bound equal to α.

H Details of Example 1

Write $X_t = \mu + \sigma Z_t$ for $t = 1, 2, \ldots$ where $Z_1, Z_2, \ldots$ are i.i.d. $\mathcal{N}(0,1)$ random variables. Substituting into the definition of S_t, we find

$$S_t = \sum_{i=1}^{t+1} \left(Z_i - \bar{Z}_{t+1}\right)^2 - t, \tag{142}$$

where $\bar{Z}_t \coloneqq t^{-1}\sum_{i=1}^t Z_i$. Evidently the distribution of S_t depends on neither μ nor σ^2. Furthermore, direct calculation shows that the increments of (S_t) may be written as

$$\Delta S_t = S_t - S_{t-1} = \frac{t}{t+1}\left(Z_{t+1} - \bar{Z}_t\right)^2 - 1 \tag{143}$$

$$=: Y_t^2 - 1, \tag{144}$$

where we define $Y_t \coloneqq \sqrt{t/(t+1)}(Z_{t+1} - \bar{Z}_t)$ for $t = 1, 2, \ldots$ (and take $S_0 = 0$ by convention). Noting that $Z_{t+1} \sim \mathcal{N}(0,1)$ is independent of $\bar{Z}_t \sim \mathcal{N}(0, t^{-1})$, we see that $Y_t \sim \mathcal{N}(0,1)$ for each t. Finally, a straightforward calculation shows that $\mathbb{E}Y_iY_j = 0$ for all $i \neq j$, so that $Y_1, Y_2, \ldots$ are i.i.d. It follows that $\Delta S_1, \Delta S_2, \ldots$ are i.i.d. centered Chi-squared random variables each with one degree of freedom. The CGF of this distribution is

$$\log \mathbb{E}e^{\lambda \Delta S_1} = -\frac{\log(1-2\lambda)}{2} - \lambda, \quad \text{for all } \lambda < \frac{1}{2}. \tag{145}$$

which is equal to $2\psi_E(\lambda)$ with scale $c = 2$. As the increments of (S_t) are i.i.d., it suffices for Definition 1 to have $\log \mathbb{E}e^{\lambda \Delta S_t} \leq \psi(\lambda)\Delta V_t$, and we have shown this holds with equality.

We have shown that (S_t) is sub-exponential with scale $c = 2$ and variance process $V_t = 2t$. Recall that Definition 1 depends only on $\lambda \geq 0$. However, since (145) holds for all $\lambda < 1/2$ and not just $0 \leq \lambda < 1/2$, replacing ΔS_t with $-\Delta S_t$ shows that $(-S_t)$ is sub-exponential with scale $c = -2$.

I Extension to smooth Banach spaces and continuous-time processes

Though we have focused on discrete-time processes taking values in $\mathbb{R}$ or $\mathcal{S}^d$, our uniform boundaries also apply to discrete-time martingales in general smooth Banach spaces and to real-valued, continuous-time martingales. In this section we briefly review concepts from Howard et al. (2020, Sections 3.4-3.5) to highlight the possibilities. First, let $(Y_t)_{t\in\mathbb{N}}$ be a martingale taking values in a separate Banach space $(\mathcal{X}, \|\cdot\|)$. Our uniform boundaries apply to any function $\Psi : \mathcal{X} \to \mathbb{R}$ satisfying the following property:

Definition 4 ((Pinelis, 1994))**.** A function $\Psi : \mathcal{X} \to \mathbb{R}$ is called $(2, D)$-*smooth* for some $D > 0$ if, for all $x, v \in \mathcal{X}$, we have (a) $\Psi(0) = 0$, (b) $|\Psi(x+v) - \Psi(x)| \leq \|v\|$, and (c) $\Psi^2(x+v) - 2\Psi^2(x) + \Psi^2(x-v) \leq 2D^2\|v\|^2$.

For example, the norm induced by the inner product in any Hilbert space is $(2,1)$-smooth, and the Schatten p-norm is $(2, \sqrt{p-1})$-smooth for $p \geq 2$.

Corollary 7. *Let $(Y_t)_{t\in\mathbb{N}}$ be a martingale taking values in a separable Banach space $(\mathcal{X}, \|\cdot\|)$, and $\Psi : \mathcal{X} \to \mathbb{R}$ is $(2, D)$-smooth; denote $D_\star \coloneqq 1 \vee D$.*

(a) Suppose $\|\Delta Y_t\| \leq c_t$ a.s. for all t for constants (c_t). Then, for any sub-Gaussian boundary f with crossing probability α and $l_0 = 2$, we have

$$\mathbb{P}\left(\exists t \geq 1 : \Psi(Y_t) \geq f\left(D_\star^2 \sum_{i=1}^t c_i^2\right)\right) \leq \alpha. \tag{146}$$

(b) Suppose $\|\Delta Y_t\| \leq c$ a.s. for all t for $c > 0$. Then, for any sub-Poisson boundary f with crossing probability α, $l_0 = 2$, and scale c, we have

$$\mathbb{P}\left(\exists t \geq 1 : \Psi(Y_t) \geq f\left(D_\star^2 \sum_{i=1}^t \mathbb{E}_{i-1}\|X_i\|^2\right)\right) \leq \alpha. \tag{147}$$

The result follows directly from the proof of Corollary 10 in Howard et al. (2020), which shows that $S_t = \Psi(Y_t)$ is sub-Gaussian or sub-Poisson with appropriate variance process (V_t) for each case, building upon the work of Pinelis (1992, 1994). For example, let (Y_t) be a martingale taking values in any Hilbert space, with $\|\cdot\|$ the induced norm, and suppose $\|\Delta Y_t\| \leq 1$ a.s. for all t. Then Corollary 7(a) with a normal mixture bound yields

$$\mathbb{P}\left(\exists t \geq 1 : \|Y_t\| \geq \sqrt{(t+\rho)\log\left(\frac{4(t+\rho)}{\alpha^2\rho}\right)}\right) \leq \alpha. \tag{148}$$

Next, let $(S_t)_{t\in\mathbb{R}_{\geq 0}}$ be a continuous-time, real-valued process. Replacing discrete-time processes in Definition 1 with continuous-time processes, and invoking the continuous-time version of Ville's inequality, our stitched, mixture and inverted stitching results extend straightforwardly to continuous time. Below we give two examples which follow from Fact 2 of Howard et al. (2020). Here $\langle S\rangle_t$ denotes the predictable quadratic variation of (S_t).

Corollary 8. *Let $(S_t)_{t\in\mathbb{R}_{\geq 0}}$ be a real-valued process.*

(a) If (S_t) is a locally square-integrable martingale with a.s. continuous paths, and f is a sub-Gaussian stitched, mixture or inverted stitching uniform boundary, then $\mathbb{P}(\exists t \in (0,\infty) : S_t \geq f(\langle S\rangle_t)) \leq e^{-2ab}$.

(b) If (S_t) is a local martingale with $\Delta S_t \leq c$ for all t, and f is a sub-Poisson mixture bound for scale c or a sub-gamma stitched bound for scale $c/3$, then $\mathbb{P}(\exists t \in (0,\infty) : S_t \geq f(\langle S\rangle_t)) \leq \alpha$.

For example, if (S_t) is a standard Brownian motion, then Corollary 8(a) with a polynomial stitched boundary yields, for any $\eta > 1, s > 1$,

$$\mathbb{P}\left(\exists t \in (0,\infty) : S_t \geq \frac{\eta^{1/4}+\eta^{-1/4}}{\sqrt{2}}\sqrt{(1\vee t)\left(s\log\log(\eta(1\vee t)) + \log\frac{\zeta(s)}{\alpha\log^s\eta}\right)}\right) \leq \alpha.$$

J Sufficient conditions for Definition 1

Table 3 offers a summary of sufficient conditions for Definition 1 to hold when (S_t) is a scalar process, while Table 4 gives conditions for matrix-valued processes. See Howard et al. (2020, Section 2) for details.

	Condition on S_t	ψ	V_t
Discrete time, $S_t = \sum_{i=1}^t X_i$*, one-sided*			
Bernoulli II	$X_t \le h, \mathbb{E}_{t-1} X_t^2 \le gh$	ψ_B	ght
Bennett	$X_t \le c$	ψ_P	$\sum_{i=1}^t \mathbb{E}_{i-1} X_i^2$
Bernstein	$\mathbb{E}_{t-1}(X_t)^k \le \frac{k!}{2} c^{k-2} \mathbb{E}_{t-1} X_t^2$	ψ_G	$\sum_{i=1}^t \mathbb{E}_{i-1} X_i^2$
*Heavy on left	$\mathbb{E}_{t-1} T_a(X_t) \le 0$ for all $a > 0$	ψ_N	$\sum_{i=1}^t X_i^2$
Bounded below	$X_t \ge -c$	ψ_E	$\sum_{i=1}^t X_i^2$
Discrete time, $S_t = \sum_{i=1}^t X_i$*, two-sided*			
Parametric	$X_t \overset{\text{iid}}{\sim} F$	$\log \mathbb{E} e^{\lambda X_1}$	t
Bernoulli I	$-g \le X_t \le h$	ψ_B	ght
Hoeffding-KS	$-g_t \le X_t \le h_t$	ψ_N	$\sum_{i=1}^t \varphi(g_i, h_i)$
Hoeffding I	$-g_t \le X_t \le h_t$	ψ_N	$\sum_{i=1}^t \left(\frac{g_i + h_i}{2}\right)^2$
*Symmetric	$X_t \sim -X_t \mid \mathcal{F}_{t-1}$	ψ_N	$\sum_{i=1}^t X_i^2$
Self-normalized I	$\mathbb{E}_{t-1} X_t^2 < \infty$	ψ_N	$\frac{1}{3} \sum_{i=1}^t \left(X_i^2 + 2\mathbb{E}_{i-1} X_i^2\right)$
Self-normalized II	$\mathbb{E}_{t-1} X_t^2 < \infty$	ψ_N	$\frac{1}{2} \sum_{i=1}^t \left((X_i)_+^2 + \mathbb{E}_{i-1}(X_i)_-^2\right)$
Cubic self-normalized	$\mathbb{E}_{t-1} \lvert X_t \rvert^3 < \infty$	ψ_G	$\sum_{i=1}^t \left(X_i^2 + \mathbb{E}_{i-1} \lvert X_i \rvert^3\right)$
Continuous time, one-sided			
Bennett	$\Delta S_t \le c$	ψ_P	$\langle S \rangle_t$
Bernstein	$W_{m,t} \le \frac{m!}{2} c^{m-2} V_t$	ψ_G	V_t
Continuous time, two-sided			
Lévy	$\mathbb{E} e^{\lambda S_1} < \infty$	$\log \mathbb{E} e^{\lambda S_1}$	t
Continuous paths	$\Delta S_t \equiv 0$	ψ_N	$\langle S \rangle_t$

Table 3: Summary of sufficient conditions a real-valued, discrete- or continuous-time process (S_t) to be sub-ψ with the given variance process. We assume (S_t) is a martingale in every case except the starred ones $(^*)$, when the first moment $\mathbb{E}\lvert X_t \rvert$ need not exist. See Howard et al. (2020, Section 2) for details. One-sided conditions yield a bound on right-tail deviations only, while two-sided conditions yield bounds on both tails. For continuous-time cases, ΔS_t denotes the jumps of (S_t) and $\langle S \rangle_t$ denotes the predictable quadratic variation. For the heavy on left case, the truncation function is defined as $T_a(y) := (y \wedge a) \vee -a$ for $a > 0$ (Bercu and Touati, 2008). The function φ used in the Hoeffding-KS case is defined in (97). The process $W_{m,t}$ in the continuous-time Bernstein case is defined in Fact 2(c) of Howard et al. (2020) (cf. van de Geer (1995)).

	Condition on $Y_t = \sum_{i=1}^t X_t$	ψ	Z_t
One-sided			
Bernoulli II	$X_t \preceq hI_d, \mathbb{E}_{t-1}X_t^2 \preceq ghI_d$	ψ_B	$ghtI_d$
Bennett	$X_t \preceq cI_d$	ψ_P	$\sum_{i=1}^t \mathbb{E}_{i-1}X_i^2$
Bernstein	$\mathbb{E}_{t-1}(X_t)^k \preceq \frac{k!}{2}c^{k-2}\mathbb{E}_{t-1}X_t^2$	ψ_G	$\sum_{i=1}^t \mathbb{E}_{i-1}X_i^2$
Bounded below	$X_t \succeq -cI_d$	ψ_E	$\sum_{i=1}^t X_i^2$
Two-sided			
Bernoulli I	$-gI_d \preceq X_t \preceq hI_d$	ψ_B	$ghtI_d$
Hoeffding-KS	$-G_tI_d \preceq X_t \preceq H_tI_d$	ψ_N	$\sum_{i=1}^t \varphi(G_i, H_i)I_d$
Hoeffding I	$-G_tI_d \preceq X_t \preceq H_tI_d$	ψ_N	$\sum_{i=1}^t \left(\frac{G_i+H_i}{2}\right)^2 I_d$
Hoeffding II	$X_t^2 \preceq A_t^2$	ψ_N	$\sum_{i=1}^t A_i^2$
*Symmetric	$X_t \sim -X_t \mid \mathcal{F}_{t-1}$	ψ_N	$\sum_{i=1}^t X_i^2$
Self-normalized I	$\mathbb{E}_{t-1}X_t^2 < \infty$	ψ_N	$\frac{1}{3}\sum_{i=1}^t \left(X_i^2 + 2\mathbb{E}_{i-1}X_i^2\right)$
Self-normalized II	$\mathbb{E}_{t-1}X_t^2 < \infty$	ψ_N	$\frac{1}{2}\sum_{i=1}^t \left((X_i)_+^2 + \mathbb{E}_{i-1}(X_i)_-^2\right)$
Cubic self-normalized	$\mathbb{E}_{t-1}\lvert X_t\rvert^3 < \infty$	ψ_G	$\sum_{i=1}^t \left(X_i^2 + \mathbb{E}_{i-1}\lvert X_i\rvert^3\right)$

Table 4: Summary of sufficient conditions for Definition 1 when $Y_t = \sum_{i=1}^t X_i$ with $X_t \in \mathcal{H}^d$, the space of Hermitian, $d \times d$ matrices, taking $S_t = \gamma_{\max}(Y_t)$ and $V_t = \gamma_{\max}(Z_t)$. We assume $\mathbb{E}X_t = 0$ and hence (Y_t) is a martingale in every case except the symmetric* case, when the first moment $\mathbb{E}|X_t|$ need not exist. See Howard et al. (2020, Section 2) for details. One-sided conditions yield a bound on right-tail deviations only, while two-sided conditions yields bounds on both tails. The function φ used in the Hoeffding-KS case is defined in (97).