The Annals of Statistics
2024, Vol. 52, No. 6, 2613–2640
https://doi.org/10.1214/24-AOS2408

TIME-UNIFORM CENTRAL LIMIT THEORY AND ASYMPTOTIC CONFIDENCE SEQUENCES

BY IAN WAUDBY-SMITH[1,a], DAVID ARBOUR[2,d], RITWIK SINHA[2,e], EDWARD H. KENNEDY[1,b] AND AADITYA RAMDAS[1,c]

[1]*Department of Statistics & Data Science, Carnegie Mellon University,* [a]*ian@ianws.com,* [b]*edward@stat.cmu.edu,* [c]*aramdas@cmu.edu*

[2]*Adobe Research, Adobe Inc.,* [d]*darbour26@gmail.com,* [e]*ritwik.sinha@gmail.com*

Confidence intervals based on the central limit theorem (CLT) are a cornerstone of classical statistics. Despite being only asymptotically valid, they are ubiquitous because they permit statistical inference under weak assumptions and can often be applied to problems even when nonasymptotic inference is impossible. This paper introduces time-uniform analogues of such asymptotic confidence intervals, adding to the literature on confidence sequences (CS)—sequences of confidence intervals that are uniformly valid over time—which provide valid inference at arbitrary stopping times and incur no penalties for "peeking" at the data, unlike classical confidence intervals which require the sample size to be fixed in advance. Existing CSs in the literature are nonasymptotic, enjoying finite-sample guarantees but not the aforementioned broad applicability of asymptotic confidence intervals. This work provides a definition for "asymptotic CSs" and a general recipe for deriving them. Asymptotic CSs forgo nonasymptotic validity for CLT-like versatility and (asymptotic) time-uniform guarantees. While the CLT approximates the distribution of a sample average by that of a Gaussian for a fixed sample size, we use strong invariance principles (stemming from the seminal 1960s work of Strassen) to uniformly approximate the entire sample average process by an implicit Gaussian process. As an illustration, we derive asymptotic CSs for the average treatment effect in observational studies (for which nonasymptotic bounds are essentially impossible to derive even in the fixed-time regime) as well as randomized experiments, enabling causal inference in sequential environments.

1. Introduction. The central limit theorem (CLT) is arguably the most widely used result in applied statistical inference, due to its ability to provide large-sample confidence intervals (CI) and p-values in a broad range of problems under weak assumptions. Examples include (a) nonparametric estimation of means, such as population proportions, (b) maximum likelihood and other M-estimation problems, and (c) modern semiparametric causal inference methodology involving (augmented) inverse propensity score weighting [6, 18, 33, 45]. Crucially, in some of these problems such as doubly robust estimation in observational studies, nonasymptotic inference is typically not possible, and hence the CLT yields asymptotic CIs for an otherwise unsolvable inference problem.

While the CLT makes efficient statistical inference possible in a broad array of problems, the resulting CIs are only valid at a prespecified sample size n, invalidating any inference that occurs at data-dependent stopping times, for example under continuous monitoring. CIs that retain validity in sequential environments are known as *confidence sequences* (CS) [7, 30] and can be used to make decisions at arbitrary stopping times (e.g., while adaptively sampling, continuously peeking at the data, etc.). CSs are an inherently nonasymptotic notion, and thus

Received July 2023; revised May 2024.
MSC2020 subject classifications. Primary 62L10, 62G20; secondary 60F17, 62D20.
Key words and phrases. Sequential, nonparametric inference, strong approximation, observational study.

essentially every published CS is nonasymptotic, including various recent state-of-the-art constructions in different settings [12, 14, 47, 50].

This paper presents a new notion: an "asymptotic confidence sequence." For the familiar reader, this might at first sound like an oxymoron. Further, it is not obvious how to posit a definition that is simultaneously sensible and tractable, meaning whether it is possible to develop such asymptotic CSs (whatever it may mean). We believe that we have formulated the "right" definition, because we accompany it with a universality result that parallels the CLT—a universal asymptotic CS that is valid under the exact same moment assumptions required by the CLT, and exploits certain time-uniform central limit theory to arrive at boundaries that one would use if the data were Gaussian. This enables the construction of asymptotic CSs in a myriad of new situations where the distributional assumptions are weak enough to remain out of the reach of nonasymptotic techniques even in fixed-time settings. The width of this universal asymptotic CS scales with the variance of the data, just like the empirical variance used in the CLT—such variance-adaptivity is only achievable for nonasymptotic methods in very specialized settings [50]. Before proceeding, let us first briefly review some notation and key facts about CSs.

1.1. *Time-uniform confidence sequences* (*CSs*). Consider the problem of estimating the population mean $\mu = \mathbb{E}(Y_1)$ from a sequence of i.i.d. data $(Y_t)_{t=1}^{\infty} \equiv (Y_1, Y_2, \dots)$ that are observed sequentially over time. A nonasymptotic $(1-\alpha)$-CI for μ is a set[1] $\dot{C}_n \equiv \dot{C}(Y_1, \dots, Y_n)$ with the property that

$$(1) \qquad \forall n \in \mathbb{N}, \quad \mathbb{P}(\mu \in \dot{C}_n) \geq 1-\alpha, \quad \text{or equivalently}, \quad \forall n \in \mathbb{N}, \quad \mathbb{P}(\mu \notin \dot{C}_n) \leq \alpha.^{2}$$

The coverage guarantee (1) of a CI is only valid at some *prespecified* sample size n, which must be decided in advance of seeing any data—peeking at the data in order to determine the sample size constitutes "p-hacking." However, it may be restrictive to fix n beforehand, and even if sample size calculations are carried out based on prior knowledge, it is impossible to know a priori whether n will be large enough to detect some signal of interest: after collecting the data, one may regret collecting too little data or much more than necessary.

CSs provide the flexibility to choose sample sizes data-adaptively while controlling the type-I error rate (see Figure 1). Formally, a CS is a sequence of CIs $(\bar{C}_t)_{t=1}^{\infty}$ such that

$$(2) \qquad \mathbb{P}(\forall t \in \mathbb{N}, \mu \in \bar{C}_t) \geq 1-\alpha, \quad \text{or equivalently}, \quad \mathbb{P}(\exists t \in \mathbb{N} : \mu \notin \bar{C}_t) \leq \alpha.$$

The statements (1) and (2) look similar but are markedly different from the data analyst's or experimenter's perspective. In particular, employing a CS has the following implications:

(a) The CS can be (optionally) updated whenever new data become available;
(b) Experiments can be continuously monitored, adaptively stopped, or continued;
(c) The type-I error is controlled at all stopping times, including data-dependent times.

In fact, CSs may be equivalently defined as CIs that are valid at arbitrary stopping times, that is,

$$\mathbb{P}(\mu \in \bar{C}_\tau) \geq 1-\alpha \quad \text{for any stopping time } \tau.$$

A proof of this equivalence can be found in Howard et al. [14], Lemma 3.

As mentioned before, while nonparametric CSs have been developed for several problems, they have thus far been *nonasymptotic*. Nonasymptotic inference for means of random variables *requires* strong assumptions on the distribution of the data [1]. These assumptions often

[1] We use overhead dots $\dot{C}_n$ to denote fixed-time (pointwise) CIs and overhead bars $\bar{C}_t$ for time-uniform CSs.

[2] Here and throughout, let $\mathbb{N} := \{1, 2, \dots\}$ and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$.

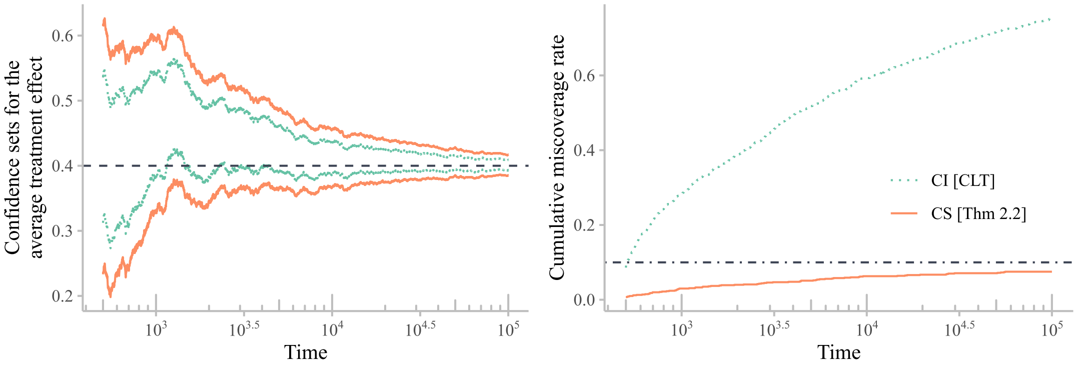

FIG. 1. *The left plot shows one run of a single experiment: an asymptotic CS alongside an asymptotic CI for a parameter of interest (in this case, the average treatment effect (ATE) of* 0.4, *an example we expand on in Section* 3). *The true value of the ATE is covered by the CS simultaneously from time* 30 *to* 10,000. *On the other hand, the CI fails to cover the true ATE at several points in time. By repeating such an experiment hundreds of times, one obtains the right plot which displays the cumulative probability of miscoverage—that is, the probability of the CS or CI failing to capture the true ATE at any time up to t. Notice that the CI error rate begins at $\alpha = 0.1$ and quickly grows, while the CS error rate never exceeds $\alpha = 0.1$.*

take the form of a parametric likelihood [12, 48], known bounds on the random variables themselves [14, 50], on their moments [47], or on their moment generating functions [14].

These added distributional assumptions make existing CSs quite unlike CLT-based CIs which (a) are universal, meaning they take the same form—up to a change in influence functions—and are computed in the same way for most problems, and (b) are often applicable even when no nonasymptotic CI is known, such as in doubly robust inference of causal effects in observational studies. Our work bridges this gap, bringing properties (a) and (b) to the anytime-valid sequential regime by making one simple modification to the usual CIs. Just as CLT-based CIs yield approximate inference for a wide variety of problems in fixed-n settings, our paper yields the same for *sequential* settings.

1.2. *Contributions and outline.* We begin by rigorously defining “asymptotic confidence sequences” (AsympCSs) in Definition 2.1 and providing a general recipe to derive explicit AsympCSs that are as easy to implement and apply as CLT-based CIs in Section 2.3. Using this recipe, we develop a Lindeberg-type AsympCS that is able to capture time-varying means under martingale dependence (Section 2.4). Furthermore, in Section 2.5, we give a definition of asymptotic time-uniform coverage (akin to coverage of asymptotic CIs) and show how *sequences* of our AsympCSs enjoy this property. In Section 3, we illustrate how the AsympCSs of Section 2.1 enable asymptotically anytime-valid semiparametric inference for causal effects in both randomized experiments and observational studies (Section 3). To be clear, we are not focused on deriving new semiparametric estimators; we simply demonstrate how semiparametric causal inference—a problem for which no known CSs exist in the observational setting—can now be tackled in fully sequential environments using the existing state-of-the-art estimators combined with our AsympCSs (Theorems 3.1 and 3.2). In Section 4, we provide a simulation study to illustrate empirical widths and miscoverage rates of AsympCSs and compare them to some existing (nonasymptotic) CSs in the literature. Finally, in Section 5 we apply the AsympCSs of Section 3 to a real observational data set by sequentially estimating the effects of fluid intake on 30-day mortality in sepsis patients. Proofs, additional results and discussions, and an R package can be found in the Supplementary Material [49]. In sum, this work expands the scope of anytime-valid inference by tackling sequential estimation problems under CLT-like moment assumptions and guarantees.

2. Asymptotic confidence sequences.

We first define what it means for a sequence of intervals to form an asymptotic confidence sequence (AsympCS). Then, we derive a "universal" AsympCS in the sense that the AsympCS does not depend on any features of the distribution beyond its mean and variance.[3] Much like classical asymptotic confidence intervals based on the CLT, this universal AsympCS fundamentally relies on Gaussian approximation. However, in this setting, the particular type of central limit theory being invoked is that of strong invariance principles, where an implicit Gaussian process is coupled with a partial sum with probability one (more details are provided in Section 2.2.2). Finally, similar to CIs based on martingale CLTs, we derive a Lindeberg-type martingale AsympCS that can track a moving average of conditional means.

2.1. *Defining asymptotic confidence sequences.*

Here, we define and present "asymptotic confidence sequences" as time-uniform analogues of CLT-based asymptotic CIs, making similarly weak moment assumptions and providing a universal closed-form boundary.

The term "asymptotic confidence sequence" may at first seem paradoxical. Indeed, ever since their introduction by Robbins and collaborators [7, 21, 22], CSs have been defined nonasymptotically, satisfying the time-uniform guarantee in equation (2). So how could a bound be both time-uniform and asymptotically valid? We clarify this critical point soon, with an analogy to classical asymptotic CIs. Similar to asymptotic CIs, AsympCSs trade nonasymptotic guarantees for (a) simplicity and universality, and (b) the ability to tackle a much wider variety of problems, especially those for which there is no known nonasymptotic CS. Said differently, AsympCSs trade finite sample validity for versatility (exemplified in Section 3 with a particular emphasis on modern causal inference).

Indeed, there is a clear desire for (asymptotically) time-uniform methods with CLT-like simplicity and versatility, especially in the context of causal inference. For example, Johari et al. [15], Section 4.3, use a Gaussian mixture sequential probability ratio test (SPRT) to conduct A/B tests (i.e. randomized experiments) for data coming from (non-Gaussian) exponential families and mentions that CLT approximations hold at large sample sizes. Similarly, Yu, Lu and Song [51] develop a mixture SPRT for causal effects in generalized linear models, where they say that their likelihood ratio forms an "approximate martingale," meaning its conditional expectation is constant up to a factor of $\exp\{o_{\mathbb{P}}(1)\}$. Moreover, Pace and Salvan [27] suggest using Robbins' Gaussian mixture CS as a closed-form "approximate CS" and they demonstrate through simulations that the time-uniform coverage guarantee tends to hold in the asymptotic regime. However, all of these approaches justify time-uniform inference with $o_{\mathbb{P}}(\cdot)$ approximations that only hold at a *fixed, pre-specified sample size*, and yet inferences are being carried out at *data-dependent sample sizes*. This section remedies the tension between fixed-n approximations and time-uniform inference by defining AsympCSs such that Gaussian approximations must hold almost surely for *all sample sizes simultaneously*. The AsympCSs we define will also be valid in a wide range of nonparametric scenarios (beyond exponential families, parametric models, and so on).

To motivate the definition of an AsympCS that follows, let us briefly review the CLT in the batch (nonsequential) setting. Suppose $Y_1, \ldots, Y_n \sim \mathbb{P}$ with mean $\mathbb{E}(Y_1) = \mu$ and variance $\operatorname{var}(Y_1) = \sigma^2$. Then the standard CLT-based CI for μ (with known variance σ) takes the form

$$\dot{C}_n := [\widehat{\mu}_n \pm \dot{\mathfrak{B}}_n] \equiv \left[\widehat{\mu}_n \pm \sigma \cdot \frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{n}}\right], \tag{3}$$

[3] We use "universal" in the same way that the CLT and law of large numbers are considered universal, as they describe macroscopic behaviors that are independent of most microscopic details of the system.

where $\widehat{\mu}_n$ is the sample mean and $\Phi^{-1}(1-\alpha/2)$ is the $(1-\alpha/2)$-quantile of a standard Gaussian $N(0,1)$ (e.g., for $\alpha = 0.05$, we have $\Phi^{-1}(0.975) \approx 1.96$). The classical notion of "asymptotic validity" is

$$\liminf_{n\to\infty} \mathbb{P}(\mu \in \dot{C}_n) \geq 1-\alpha.$$

While the above is the standard definition of an asymptotic CI, one could have arrived at an alternative definition by noting the following rather strong statement that can be made under the same conditions: there exist[4] i.i.d. standard Gaussians $Z_1, \dots, Z_n \sim N(0,1)$ such that

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu)/\sigma = \frac{1}{n}\sum_{i=1}^{n} Z_i + o_{\mathbb{P}}(1/\sqrt{n}). \tag{4}$$

From this vantage point, we note that the CI in (3) has the additional guarantee that there in fact exists an (unknown) *nonasymptotic* $(1-\alpha)$-CI $[\widehat{\mu}_n \pm \dot{\mathfrak{B}}_n^\star]$ such that

$$\dot{\mathfrak{B}}_n^\star / \dot{\mathfrak{B}}_n \xrightarrow{\mathbb{P}} 1. \tag{5}$$

We deliberately highlight the above property of asymptotic CIs because it ends up serving as a natural starting point for defining asymptotic confidence *sequences*. In particular, we will define AsympCSs so that an analogous approximation to (5) holds uniformly over time, almost surely. Statements like (4) are known as "couplings" and appear in the literature on strong approximations and invariance principles where similar guarantees can indeed be shown to hold almost surely and at faster rates under additional moment assumptions [10, 19, 20].

DEFINITION 2.1 (Asymptotic confidence sequences). Let $\mathcal{T}$ be a totally ordered infinite set (denoting time) that has a minimum value $t_0 \in \mathcal{T}$. We say that the intervals $(\widehat{\theta}_t - L_t, \widehat{\theta}_t + U_t)_{t\in\mathcal{T}}$ centered at the estimators $(\widehat{\theta}_t)_{t\in\mathcal{T}}$ with nonzero bounds $L_t, U_t > 0$, $\forall t \in \mathcal{T}$ form a $(1-\alpha)$-*asymptotic confidence sequence* (AsympCS) for a sequence of real parameters $(\theta_t)_{t\in\mathcal{T}}$ if there exists a (typically unknown) nonasymptotic $(1-\alpha)$-CS $(\widehat{\theta}_t - L_t^\star, \widehat{\theta}_t + U_t^\star)_{t\in\mathcal{T}}$ for $(\theta_t)_{t\in\mathcal{T}}$—that is, satisfying

$$\mathbb{P}\bigl(\forall t \in \mathcal{T}, \theta_t \in \bigl[\widehat{\theta}_t - L_t^\star, \widehat{\theta}_t + U_t^\star\bigr]\bigr) \geq 1-\alpha,$$

and such that L_t, U_t become arbitrarily precise almost-sure approximations to $L_t^\star$ and $U_t^\star$:

$$L_t^\star / L_t \xrightarrow{\text{a.s.}} 1 \quad \text{and} \quad U_t^\star / U_t \xrightarrow{\text{a.s.}} 1.$$

In words, Definition 2.1 says that an AsympCS $(C_t)_{t\in\mathcal{T}}$ centered at $(\widehat{\theta}_t)_{t\in\mathcal{T}}$ is an arbitrarily precise approximation of some nonasymptotic CS $(C_t^\star)_{t\in\mathcal{T}}$ centered at $(\widehat{\theta}_t)_{t\in\mathcal{T}}$ as $t \to \infty$. Throughout the paper, we will mostly focus on the case where $\mathcal{T} = \mathbb{N}_0$ with $t_0 = 0$.

It is important to note that alternate definitions fail to be coherent in different ways. As one example, we could have hypothetically defined a sequence of intervals $(C_t(\alpha))_{t\in\mathcal{T}}$ to be a $(1-\alpha)$-AsympCS if $\limsup_{m\to\infty} \mathbb{P}(\exists t \geq m : \mu \notin C_t(\alpha)) \leq \alpha$, analogously to asymptotic CIs which satisfy $\limsup_{n\to\infty} \mathbb{P}(\mu \notin \dot{C}_n(\alpha)) \leq \alpha$. In words, we could have posited that if we just start peeking late enough, then the probability of eventual miscoverage would indeed be below α. Unfortunately, even for nonasymptotic CSs constructed at any level $\alpha' \in (0,1)$, the former limit is *zero*, so this inequality would be vacuously true, regardless of α', even if $\alpha' \gg \alpha$. However, we do show that *sequences* of our AsympCSs satisfy a guarantee of this type (see Section 2.5), but we delay those definitions until later as they are more involved.

[4] Technically, writing (4) may require enriching the probability space so that both Y and Z can be defined (but without changing their laws). See Einmahl [10], Equation (1.2), for a precise statement.

By virtue of being defined in terms of their limiting behavior, one can obviously construct AsympCSs (as well as asymptotic confidence intervals) with nonsensical finite-sample behavior. It is thus imperative that if a practitioner decides to employ an asymptotic method, they do so with the understanding that its effectiveness relies on it exploiting some well-approximated nonasymptotic phenomenon, and that its limiting behavior should be viewed as a rigorous manifestation of a guiding principle rather than a panacea.

REMARK 1 (Why almost surely?). One may wonder why it is necessary to define AsympCSs so that $L_t^\star / L_t \to 1$ *almost surely* (rather than in probability, for example). Since CSs are bounds that hold uniformly over time with high probability, convergence in probability $L_t^\star / L_t = 1 + o_{\mathbb{P}}(1)$ is not the right notion of convergence, as it only requires that the approximation term $o_{\mathbb{P}}(1)$ be small with high probability for sufficiently large *fixed* t, but not for all t uniformly. It is natural to try to extend convergence in probability to *time-uniform convergence with high probability*—that is, $\sup_{k \geq t}(L_k^\star / L_k) = 1 + o_{\mathbb{P}}(1)$—but it turns out that this is in fact *equivalent* to almost-sure convergence $L_t^\star / L_t = 1 + o_{\text{a.s.}}(1)$; see Appendix B.3.

Going forward, we may omit "a.s." from $o_{\text{a.s.}}(\cdot)$ and $O_{\text{a.s.}}(\cdot)$ and instead simply write $o(\cdot)$ and $O(\cdot)$, respectively to simplify notation. Now that we have defined AsympCSs as time-uniform analogues of asymptotic CIs, we will explicitly derive AsympCSs for the mean of i.i.d. random variables with finite variances (i.e., under the same assumptions as the CLT).

2.2. *Warmup*: *AsympCSs for the mean of i.i.d. random variables.* We now construct an explicit AsympCS for the mean of i.i.d. random variables by combining a variant of Robbins' (nonasymptotic) Gaussian mixture boundary [30] with Strassen's strong approximation theorem [39]. Before presenting the result, let us review Robbins' boundary and Strassen's result, and discuss how they can be used in conjunction to arrive at the AsympCS in Theorem 2.2.

2.2.1. *Robbins' Gaussian mixture boundary.* The study of CSs began with Herbert Robbins and colleagues [7, 21, 22, 30, 31], leading to several fundamental results and techniques including the famous Gaussian mixture boundary for partial sums of i.i.d. Gaussian random variables [30] (see also Howard et al. [14], §3.2) which we recall here. Suppose $(Z_t)_{t=1}^{\infty}$ are i.i.d. standard Gaussian random variables. Then for any $\rho > 0$,

$$\mathbb{P}\left(\exists t \geq 1 : \left|\frac{1}{t}\sum_{i=1}^{t} Z_i\right| \geq \sqrt{\frac{t\rho^2 + 1}{t^2\rho^2}\log\left(\frac{t\rho^2+1}{\alpha^2}\right)}\right) \leq \alpha. \tag{6}$$

Notice that the above boundary scales as $O(\sqrt{\log t / t})$ for any $\rho > 0$. In fact, Robbins [30], equation 11, noted that (6) holds not only for Gaussian random variables, but for those that are 1-sub-Gaussian, and hence pre-multiplying the boundary by σ yields a σ-sub-Gaussian time-uniform concentration inequality, serving as a time-uniform analogue of Chernoff or Hoeffding inequalities. The connections between these fixed-time and time-uniform concentration inequalities are made explicit in Howard et al. [14]. Nevertheless, Eg. (6) requires *a priori* knowledge of $\sigma > 0$ unlike CLT-based CIs which we aim to emulate in the (asymptotically) time-uniform regime. The following strong Gaussian approximation due to Strassen [39] will serve as a technical tool allowing the nonasymptotic sub-Gaussian bound in (6) to be applied to partial sums of arbitrary random variables with finite variances.

2.2.2. *Strassen's strong approximation.* Strassen [39] initiated the study of "strong approximation" (also called strong invariance principles or strong embeddings) which blossomed into an active and impactful corner of probability theory research over the subsequent years, culminating in now-classical results such as the Komlós–Major–Tusnády embeddings [19, 20, 24] and other related works [5, 26, 40].

In his 1964 paper, Strassen [39], §2, used the Skorokhod embedding [38] (see also [3], p. 513) to obtain a strong invariance principle which connects asymptotic Gaussian behavior with the law of the iterated logarithm. Concretely, let $(Y_t)_{t=1}^\infty$ be an infinite sequence of i.i.d. random variables from a distribution $\mathbb{P}$ with mean μ and variance σ^2. Then, on a potentially richer probability space,[5] there exist standard Gaussian random variables $(Z_t)_{t=1}^\infty$ whose partial sums are almost-surely coupled with those of $(Y_t)_{t=1}^\infty$ up to iterated logarithm rates, that is,

$$\left|\sum_{i=1}^t (Y_i - \mu)/\sigma - \sum_{i=1}^t Z_i\right| = o(\sqrt{t \log\log t}) \quad \text{almost surely.} \tag{7}$$

Notice that the law of the iterated logarithm states that $|\sum_{i=1}^t (Y_i - \mu)/\sigma| = O(\sqrt{t \log\log t})$ while (7) states that the same partial sum is almost-surely coupled with an implicit Gaussian process—that is, replacing $O(\cdot)$ with $o(\cdot)$. It may be convenient to divide by t and interpret (7) on the level of sample averages rather than partial sums, in which case the right-hand side becomes $o(\sqrt{\log\log t/t})$. Let us now describe how Strassen's strong approximation can be combined with Robbins' Gaussian mixture boundary to derive an AsympCS under finite moment assumptions akin to the CLT.

2.2.3. *The Gaussian mixture asymptotic confidence sequence.*

Given the juxtaposition of (6) and (7), the high-level approach to the derivation of AsympCSs becomes clearer. Indeed, the essential idea behind Theorem 2.2 is as follows. By Strassen's strong approximation, we couple the partial sums $S_t := \sum_{i=1}^t (Y_i - \mu)/\sigma$ with implicit partial sums $G_t := \sum_{i=1}^t Z_i$ of Gaussians $(Z_t)_{t=1}^\infty$, and then use Robbins' mixture boundary to obtain a time-uniform high-probability bound on the deviations of $|G_t|$, noting that the coupling rate $o(\sqrt{t \log\log t})$ is asymptotically dominated by the concentration rate $O(\sqrt{t \log t})$, leading to asymptotic validity in the formal sense of Definition 2.1.

THEOREM 2.2 (Gaussian mixture asymptotic confidence sequence). *Suppose $(Y_t)_{t=1}^\infty \sim \mathbb{P}$ is an infinite sequence of i.i.d. observations from a distribution $\mathbb{P}$ with mean μ and finite variance. Let $\widehat{\mu}_t := \frac{1}{t}\sum_{i=1}^t Y_i$ be the sample mean, and $\widehat{\sigma}_t^2 := \frac{1}{t}\sum_{i=1}^t Y_i^2 - (\widehat{\mu}_t)^2$ the sample variance based on the first t observations. Then, for any prespecified constant $\rho > 0$,*

$$\overline{C}_t^{\mathcal{G}} \equiv (\widehat{\mu}_t \pm \overline{\mathfrak{B}}_t^{\mathcal{G}}) := \left(\widehat{\mu}_t \pm \widehat{\sigma}_t \sqrt{\frac{t\rho^2 + 1}{t^2\rho^2} \log\left(\frac{t\rho^2+1}{\alpha^2}\right)}\right) \tag{8}$$

forms a $(1-\alpha)$-AsympCS for μ.

The proof of Theorem 2.2 is in Appendix A.1. We can think of $\rho > 0$ as a user-chosen tuning parameter which dictates the time at which (8) is tightest, and we discuss how to easily tune this value in Appendix B.2. A one-sided analogue of (8) can be found in Appendix B.1.

While (8) may look visually similar to Robbins' (sub)-Gaussian mixture CS [30]—written explicitly in Howard et al. [14], equation (14),—it is worth pausing to reflect on how they are markedly different. First, Robbins' CS is a nonasymptotic bound that is only valid for

[5] A richer probability space may be needed to describe Gaussian random variables, if for example, $(Y_t)_{t=1}^\infty$ are $\{0, 1\}$-valued on a probability space whose probability measure is dominated by the counting measure. This construction of a richer probability space imposes no additional assumptions on $(Y_t)_{t=1}^\infty$, and is only a technical device used to rigorously couple two sequences of random variables, and appears in essentially all papers on strong invariance principles, not just Strassen [39].

σ-sub-Gaussian random variables, meaning $\mathbb{E}\exp\{\lambda(Y_1 - \mathbb{E}Y_1)\} \le \exp\{\sigma^2\lambda^2/2\}$ for some *a priori* known $\sigma > 0$, while Theorem 2.2 does not require the existence of a finite MGF at all (much less a known upper bound on it). Second, Robbins' CS uses this known (possibly conservative) σ in place of $\widehat{\sigma}_t$ in (8), and thus it cannot adapt to an unknown variance, while (8) always scales with $\sqrt{\mathrm{var}(Y_1)}$. In simpler terms, Theorem 2.2 is an asymptotically time-uniform analogue of the CLT in the same way that Robbins' CS is a time-uniform analogue of a sub-Gaussian concentration inequality (e.g., Hoeffding's inequality [11, 13]).

It is important not to confuse Theorem 2.2 with a martingale CLT as the latter still gives fixed-time CIs in the spirit of the usual CLT but under different assumptions on the martingale difference sequence (however, we do present an analogue of Theorem 2.2 under martingale dependence in Proposition 2.5).

2.2.4. *An asymptotic confidence sequence with iterated logarithm rates.* As a consequence of the law of the iterated logarithm, a confidence sequence for μ centered at $\widehat{\mu}_t$ cannot have an asymptotic width smaller than $O(\sqrt{\log\log t/t})$. This is easy to see since

$$\limsup_{t\to\infty} \frac{\sqrt{t}|\widehat{\mu}_t - \mu|}{\sigma\sqrt{2\log\log t}} = 1.$$

This raises the question as to whether $\overline{C}_t^{\mathcal{G}}$ can be improved so that the optimal asymptotic width of $O(\sqrt{\log\log t/t})$ is achieved. Indeed, we can replace Robbins' Gaussian mixture boundary with Howard et al. [14], equation (2), (or virtually any other Gaussian boundary for that matter) in the proof of Theorem 2.2 to derive such an AsympCS, but as the authors discuss, mixture boundaries such as the one in Theorem 2.2 may be preferable in practice, because any bound that is tighter "later on" (asymptotically) must be looser "early on" (at practical sample sizes) due to the fact that all such bounds have a cumulative miscoverage probability $\le \alpha$. This is formally a concern for nonasymptotic CSs, but only applies to AsympCSs insofar as they are asymptotic approximations of nonasymptotic bounds. Nevertheless, we present an AsympCS with an iterated logarithm rate here for completeness.

PROPOSITION 2.3 (Iterated logarithm asymptotic confidence sequences). *Under the same conditions as Theorem 2.2,*

$$\overline{C}_t^{\mathcal{L}} \equiv (\widehat{\mu}_t \pm \overline{\mathfrak{B}}_t^{\mathcal{L}}) := \left(\widehat{\mu}_t \pm \widehat{\sigma}_t \cdot 1.7\sqrt{\frac{\log\log(2t) + 0.72\log(10.4/\alpha)}{t}}\right)$$

forms a $(1-\alpha)$*-AsympCS for* μ.

We omit the proof of Proposition 2.3 as it proceeds in a similar fashion to that of Theorem 2.2. In fact, both of these AsympCSs are simply instantiations of a more general recipe for deriving AsympCSs by combining strong approximations with time-uniform boundaries for the approximating process, an approach that we discuss in the following section.

2.3. *A general recipe for deriving asymptotic confidence sequences.* The proofs of both Theorem 2.2 and Proposition 2.3 follow the same general structure, combining strong approximations with time-uniform boundaries along with some other almost-sure asymptotic behavior. Abstracting away the details specific to these particular results, we provide the following four general conditions under which many AsympCSs can be derived, including those from the previous section but also Lyapunov- and Lindeberg-type AsympCS that we will state in Section 2.4).

In what follows, let $\mathcal{T}$ be a totally ordered infinite set that includes a minimum value $t_0 \in \mathcal{T}$ (for example, one may think about $\mathcal{T}$ as $\mathbb{R}^{\ge 0}$ or $\mathbb{N}_0$ with $t_0 = 0$) and let $(\widehat{\theta}_t)_{t\in\mathcal{T}}$ be a sequence of estimators for the real-valued parameters $(\theta_t)_{t \in \mathcal{T}}$. Then, consider the following four conditions where we use the "Condition G-X" enumeration as a mnemonic for the X$^{\text{th}}$ condition in the section on General recipes for AsympCSs).

CONDITION G-1 (Strong approximation). *On a potentially enriched probability space, there exists a process $(Z_t)_{t \in \mathcal{T}}$ starting at $Z_{t_0} \equiv 0$ that strongly approximates $(\theta_t - \widehat{\theta}_t)_{t \in \mathcal{T}}$ up to a rate of $(r_t)_{t \in \mathcal{T}}$, i.e.*

$$(\theta_t - \widehat{\theta}_t) - Z_t = O(r_t) \quad \textit{almost surely}. \tag{9}$$

CONDITION G-2 (Boundary for the approximating process). *There exist $\widehat{L}_t > 0$ and $\widehat{U}_t > 0$ for each $t \in \mathcal{T}$ so that $[-\widehat{L}_t, \widehat{U}_t]_{t \in \mathcal{T}}$ forms a $(1-\alpha)$-boundary for the process $(Z_t)_{t \in \mathcal{T}}$ given in (9):*

$$\mathbb{P}(\forall t \in \mathcal{T}, Z_t \in [-\widehat{L}_t, \widehat{U}_t]) \geq 1 - \alpha. \tag{10}$$

CONDITION G-3 (Strong approximation rate). *The approximation rate $(r_t)_{t \in \mathcal{T}}$ in (9) is faster than both $(\widehat{L}_t)_{t \in \mathcal{T}}$ and $(\widehat{U}_t)_{t \in \mathcal{T}}$ in (10), that is,*

$$r_t = o(\widehat{L}_t \wedge \widehat{U}_t) \quad \textit{almost surely.}$$

CONDITION G-4 (Almost-sure approximate boundary). *The $(1-\alpha)$-boundary $[-\widehat{L}_t, \widehat{U}_t]_{t \in \mathcal{T}}$ for $(Z_t)_{t \in \mathcal{T}}$ is almost-surely approximated by the sequence $[-L_t, U_t]_{t \in \mathcal{T}}$, that is,*

$$L_t / \widehat{L}_t \xrightarrow{\text{a.s.}} 1 \quad \textit{and} \quad U_t / \widehat{U}_t \xrightarrow{\text{a.s.}} 1.$$

Deriving new AsympCSs then reduces to the conceptually simpler but nevertheless nontrivial task of satisfying the requisite conditions above. For example, in the previous section, we satisfied Condition G-1 via Strassen [39], Condition G-2 via Robbins [30], Condition G-3 via the combination of Strassen [39] and Robbins [30], and Condition G-4 via the strong law of large numbers (SLLN). The only difference between Theorem 2.2 and Proposition 2.3 was in what boundaries were being used for $[L_t, U_t]_{t \in \mathcal{T}}$ and $[\widehat{L}_t, \widehat{U}_t]_{t \in \mathcal{T}}$. More generally, under Conditions G-1–G-4, we have the following abstract theorem for AsympCSs.

THEOREM 2.4 (An abstract AsympCS for well-approximated processes). *Let $\mathcal{T}$ be a totally ordered infinite set containing a minimal element $t_0 \in \mathcal{T}$ and let $(\widehat{\theta}_t)_{t \in \mathcal{T}}$ be a real-valued process. Under Conditions G-1–G-4,*

$$[\widehat{\theta}_t - L_t, \widehat{\theta}_t + U_t]$$

forms a $(1-\alpha)$-AsympCS for θ_t meaning there exists (on a potentially enriched probability space) some nonasymptotic $(1-\alpha)$-CS $[\widehat{\theta}_t - L_t^\star, \widehat{\theta}_t + U_t^\star]_{t \in \mathcal{T}}$ for $(\theta_t)_{t \in \mathcal{T}}$, that is,

$$\mathbb{P}(\forall t \in \mathcal{T}, \theta_t \in [\widehat{\theta}_t - L_t^\star, \widehat{\theta}_t + U_t^\star]) \geq 1 - \alpha$$

such that

$$L_t^\star / L_t \xrightarrow{\text{a.s.}} 1 \quad \textit{and} \quad U_t^\star / U_t \xrightarrow{\text{a.s.}} 1.$$

We provide a short proof of Theorem 2.4 in Appendix A.2. Note that the lower boundaries given by $L_t^\star$ and $\widehat{L}_t$ are not the same, but rather $L_t^\star$ is constructed from $\widehat{L}_t$ (and similarly for $U_t^\star$ and $\widehat{U}_t$). In the following section, we will use the general recipe of Theorem 2.4 to obtain AsympCSs for time-varying means from non-i.i.d. random variables under martingale dependence akin to the Lindeberg CLT [3, 23].

2.4. *Lindeberg- and Lyapunov-type AsympCSs for time-varying means.* The results in Theorem 2.2 and Proposition 2.3 focused on the situation where the observed random variables are i.i.d. as this is one of the most commonly studied regimes in statistical inference. One may also be interested in the case where means and variances do not remain constant over time, or where observations are dependent. We will now show that an analogue of Theorem 2.2 holds for random variables with time-varying *means and variances* under *martingale dependence*. In this case, rather than the AsympCS covering some fixed μ, it covers the average conditional mean thus far: $\widetilde{\mu}_t := \frac{1}{t}\sum_{i=1}^t \mu_i$—to be made precise shortly.[6]

Given the additional complexity introduced by considering time-varying conditional distributions, we will first explicitly spell out the conditions required to achieve a time-varying analogue of Theorem 2.2. Suppose $(Y_t)_{t=1}^\infty$ is a sequence of random variables with conditional means and variances given by $\mu_t := \mathbb{E}(Y_t | Y_1^{t-1})$ and $\sigma_t^2 := \mathrm{var}(Y_t | Y_1^{t-1})$, respectively where we use the shorthand Y_1^{t-1} for $\{Y_1, \ldots, Y_{t-1}\}$. First, we require that the average conditional variance $\widetilde{\sigma}_t^2 := \frac{1}{t}\sum_{i=1}^t \sigma_i^2$ either does not vanish, or does so superlinearly; equivalently, we require that the cumulative conditional variance diverges almost surely.

CONDITION L-1 (Cumulative variance diverges almost surely). *For each $t \geq 1$, let $\sigma_t^2 :=$ $\mathrm{var}(Y_t | Y_1^{t-1})$ be the conditional variance of Y_t. Then*

$$V_t := \sum_{i=1}^t \sigma_i^2 \to \infty \quad \textit{almost surely.}$$

Condition L-1 can also be interpreted as saying that the average conditional variance $\widetilde{\sigma}_t^2 := \frac{1}{t}\sum_{i=1}^t \sigma_i^2$ does not vanish faster than $1/t$ (if at all), meaning $\widetilde{\sigma}_t^2 = \omega_{\text{a.s.}}(1/t)$. For example, Condition L-1 would hold if $\widetilde{\sigma}_t^2 \xrightarrow{\text{a.s.}} \sigma_\star^2$ for some $\sigma_\star^2 > 0$ or in the i.i.d. case where $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_\star^2$. Second, we require a Lindeberg-type uniform integrability condition on the tail behavior of $(Y_t)_{t=1}^\infty$.

CONDITION L-2 (Lindeberg-type uniform integrability). *There exists some $0 < \kappa < 1$ such that*

$$\sum_{t=1}^\infty \frac{\mathbb{E}[(Y_t - \mu_t)^2 \mathbb{1}((Y_t - \mu_t)^2 > V_t^\kappa) | Y_1^{t-1}]}{V_t^\kappa} < \infty \quad \textit{almost surely.}$$

Notice that Condition L-2 is satisfied if all conditional q^{th} moments are almost surely uniformly bounded for some $q > 2$, meaning $1/K \leq \mathbb{E}(|Y_t - \mu_t|^q | Y_1^{t-1}) < K$ a.s. for all $t \geq 1$ and for some constant $K > 0$, or more generally under a Lyapunov-type condition that states $\sum_{t=1}^\infty [\mathbb{E}(|Y_t - \mu_t|^{2+\delta} | Y_1^{t-1}) / \sqrt{V_t}^{2+\delta}] < \infty$ a.s. for some $\delta > 0$.[7] Third and finally, we require a consistent variance estimator.

CONDITION L-3 (Consistent variance estimation). *Let $\widehat{\sigma}_t^2$ be an estimator of $\widetilde{\sigma}_t^2$ constructed using $Y_1, \ldots, Y_t$ such that*

$$\widehat{\sigma}_t^2 / \widetilde{\sigma}_t^2 \xrightarrow{\text{a.s.}} 1. \tag{11}$$

[6] Throughout this section and the remainder of the paper, we use the overhead tilde (e.g., $\widetilde{\mu}_t$, $\widetilde{\sigma}_t$, and $\widetilde{C}_t$) to emphasize that these quantities can change over time. For example, Figure 2 explicitly displays means and CSs with sinusoidal behaviors resembling a tilde.

[7] We show that the Lyapunov-type condition implies Condition L-2 in Appendix B.5.

Note that in the i.i.d. case, (11) would hold using the sample variance by the SLLN. More generally for independent but non-identically distributed data, Condition L-3 holds as long as the variation in means vanishes—that is, $\frac{1}{t}\sum_{i=1}^{t}(\mu_i - \widetilde{\mu}_t)^2 = o(1)$—but we will expand on this later in Corollary 2.6. Given Conditions L-1, L-2, and L-3, we have the following AsympCS for the time-varying conditional mean $\widetilde{\mu}_t := \frac{1}{t}\sum_{i=1}^{t}\mu_i$.

PROPOSITION 2.5 (Lindeberg-type Gaussian mixture martingale AsympCS). *Let* $(Y_t)_{t=1}^{\infty}$ *be a sequence of random variables with conditional mean* $\mu_t := \mathbb{E}(Y_t | Y_1^{t-1})$ *and conditional variance* $\sigma_t^2 := \operatorname{var}(Y_t | Y_1^{t-1})$. *Then under Assumptions* L-1, L-2, *and* L-3, *we have that*

$$\widetilde{C}_t \equiv (\widehat{\mu}_t \pm \widetilde{\mathfrak{B}}_t) := \left(\widehat{\mu}_t \pm \sqrt{\frac{t\widehat{\sigma}_t^2\rho^2 + 1}{t^2\rho^2}\log\left(\frac{t\widehat{\sigma}_t^2\rho^2 + 1}{\alpha^2}\right)}\right)$$

forms a $(1-\alpha)$-*AsympCS for the running average conditional mean* $\widetilde{\mu}_t := \frac{1}{t}\sum_{i=1}^{t}\mu_i$.

At a high level, the proof of Proposition 2.5 (found in Appendix A.3) follows from the general AsympCS procedure of Theorem 2.4 by using Condition L-2 and Strassen's 1967 strong approximation [40] (not to be confused with his 1964 [39] result that we used in Theorem 2.2) to satisfy Conditions G-1 and G-3, and a variant of Robbins' mixture martingale for non-i.i.d. random variables along with Conditions L-1 and L-3 to satisfy Conditions G-2 and G-4.

Notice that if the data happen to be i.i.d., then $(\widetilde{C}_t)_{t=1}^{\infty}$ is asymptotically equivalent to $(\overline{C}_t^{\mathcal{G}})_{t=1}^{\infty}$ given in Theorem 2.2 (here, "asymptotic equivalence" simply means that the ratio of the two boundaries converges a.s. to 1). In other words, Proposition 2.5 is valid in a more general (non-i.i.d.) setting, but will essentially recover Theorem 2.2 in the i.i.d. case. Figure 2 illustrates what $\widetilde{C}_t$ may look like in practice. Note that when $(Y_t)_{t=1}^{\infty}$ are independent with $\mu_1 = \mu_2 = \cdots = \mu_\star$, and $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_\star^2$, it is nevertheless the case that $\widetilde{C}_t$ forms a $(1-\alpha)$-AsympCS for $\mu_\star$ under the same assumptions as Theorem 2.2. In this sense, we can view $(\widetilde{C}_t)_{t=1}^{\infty}$ as "robust" to deviations from independence and stationarity.[8] A one-sided analogue of Proposition 2.5 is presented in Proposition B.2 within Appendix B.1.

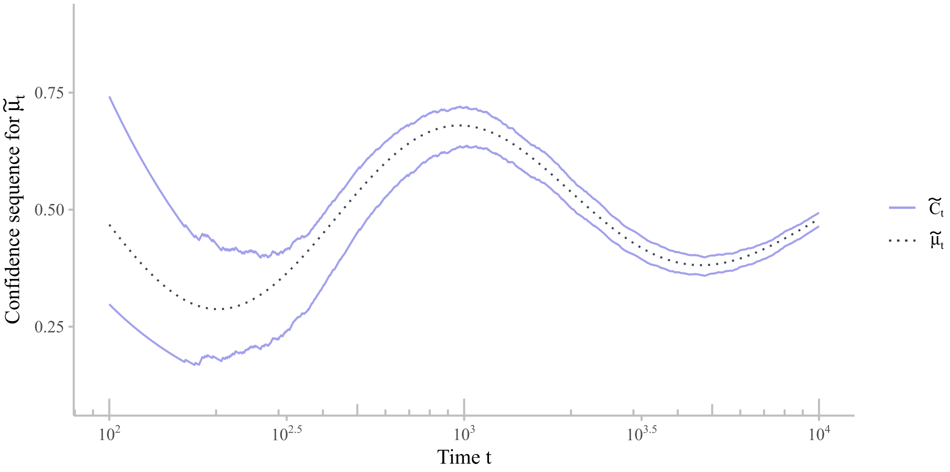

FIG. 2. *A 90%-AsympCS for the time-varying mean* $\widetilde{\mu}_t$ *using Proposition* 2.5 *with* ρ *optimized for* $t^\star = 500$ *based on the exact solution of Appendix B.2. Here, we have set* $\mu_t := \frac{1}{2}(1 - \sin(2\log(e + 10t))/\log(e + 0.01t))$ *to produce the sinusoidal behavior of* $\widetilde{\mu}_t$. *Notice that* $\widetilde{C}_t$ *uniformly captures* $\widetilde{\mu}_t$, *adapting to its nonstationarity.*

[8] Here, the term "robust" should not be interpreted in the same spirit as "doubly robust," where the latter is specific to the discussions surrounding functional estimation and causal inference in Section 3.

As a near-immediate corollary of Proposition 2.5, we have the following Lyapunov-type AsympCS under independent but non-identically distributed random variables.

COROLLARY 2.6 (Lyapunov-type AsympCS). *Suppose $(Y_t)_{t=1}^{\infty}$ is a sequence of independent random variables with individual means and variances given by $\mu_t := \mathbb{E}(Y_t)$ and $\sigma_t^2 := \mathrm{var}(Y_t)$, respectively. Suppose that in addition to Condition L-1 and the Lyapunov-type condition $\sum_{i=1}^{\infty}[\mathbb{E}|Y_i - \mu_i|^{2+\delta}/\sqrt{V_i}^{2+\delta}] < \infty$, we have the following regularity conditions*:

$$\sum_{i=1}^{\infty} \frac{\mathbb{E}|Y_i^2 - \mathbb{E}Y_i^2|^{1+\beta}}{V_i^{1+\beta}} < \infty, \qquad \widetilde{\mu}_t^2 = o(V_t) \quad \text{a.s.}, \quad \text{and} \quad \frac{1}{t}\sum_{i=1}^{t}(\mu_i - \widetilde{\mu}_t)^2 = o(\widetilde{\sigma}_t^2) \tag{12}$$

for some $\beta \in (0, 1)$. In other words, the higher moments of $(Y_t)_{t=1}^{\infty}$, the running mean $\widetilde{\mu}_t$, and the cumulative "variation in means" $\sum_{i=1}^{t}(\mu_i - \widetilde{\mu}_t)^2$ all cannot diverge too quickly relative to $(V_t)_{t=1}^{\infty}$. Then using the sample variance for $\widehat{\sigma}_t^2$, $\widetilde{C}_t$ forms a $(1-\alpha)$-AsympCS for the running average mean $\widetilde{\mu}_t := \frac{1}{t}\sum_{i=1}^{t}\mu_i$.

Clearly, the conditions of (12) are trivially satisfied if the $(2 + 2\beta)^{\text{th}}$ absolute central moments are uniformly bounded over time and if the means converge. Since Conditions L-1 and L-2 hold by the assumptions of Corollary 2.6, the proof in Appendix A.4 simply shows how the conditions in (12) imply Condition L-3.

As suggested by Section 2.3, we can combine Theorem 2.4 with essentially any other Gaussian boundary, and indeed there are others that can yield Lindeberg- and Lyapunov-type AsympCSs but we do not enumerate any more here, though we do mention one inspired by Robbins [30], equation (20), in passing in Section 2.6. The next section discusses how all of the aforementioned AsympCSs satisfy a certain formal asymptotic coverage guarantee.

2.5. *Asymptotic coverage and type-I error control.* While the AsympCSs derived thus far serve as sequential analogues of CLT-based CIs, it is not immediately obvious whether the bounds introduced in the previous section enjoy similar asymptotic coverage (equivalently, type-I error) guarantees. We will now give a positive answer to this question by showing that after appropriate tuning, our AsympCSs have asymptotic $(1-\alpha)$-coverage uniformly for all $t \geq m$ as $m \to \infty$ (to be formalized in Definition 2.7).

Recall that the coverage of CLT-based CIs is at least $(1-\alpha)$ in the limit:

$$\liminf_{n\to\infty} \mathbb{P}(\mu \in \dot{C}_n) \geq 1 - \alpha, \tag{13}$$

but what is the right time-uniform analogue of (13)? Since any single AsympCS will simply have *some* coverage, we provide the following definition as a time-uniform analogue of (13) for *sequences* of sets that start later and later. In the definition that follows, $m \geq 1$ will play the role of an "initial peeking time" and time-uniformity will be provided with respect to all $t \geq m$.

DEFINITION 2.7 (Asymptotic time-uniform coverage). For each $m \in \mathbb{N}$, let $(C_t(m))_{t=m}^{\infty}$ be a sequence of sets, and let $\alpha \in (0, 1)$ be the desired miscoverage level. We say that $(C_t(m))_{t=m}^{\infty}$ has *asymptotic time-uniform* $(1-\alpha)$*-coverage* for $(\mu_t)_{t=1}^{\infty}$ if

$$\liminf_{m\to\infty} \mathbb{P}\big(\forall t \geq m, \mu_t \in C_t(m)\big) \geq 1 - \alpha, \tag{14}$$

and we say that this coverage is *sharp* if the above inequality holds with equality and the limit infimum is replaced by a limit.

To the best of our knowledge, the existing literature lacks a concrete definition of asymptotic time-uniform coverage (or type-I error control) like Definition 2.7, but sequences of AsympCSs satisfying (14) have been implicit in Robbins [30] and Robbins and Siegmund [31], and the followup work of Bibaut, Kallus and Lindon [2]. In what follows, we provide (sharp) coverage guarantees for our AsympCSs. Furthermore, in Section 2.6 we strictly improve on aforementioned bounds by Robbins [30] and Robbins and Siegmund [31]. Furthermore, we note that a bound Bibaut, Kallus and Lindon [2] is in a certain sense equivalent to one that we provide here.

In order to obtain asymptotic time-uniform coverage, we need a stronger variant of Condition L-3 so that variances are estimated at polynomial rates (rather than at arbitrary rates).

CONDITION L-3-η (Polynomial rate variance estimation). *There exists some* $0 < \eta < 1$ *such that*

$$\widehat{\sigma}_t^2 - \widetilde{\sigma}_t^2 = o\left(\frac{(t\widetilde{\sigma}_t^2)^\eta}{t}\right) \quad \textit{almost surely}. \tag{15}$$

Note that while Condition L-3-η is stronger than Condition L-3, it is still quite mild. For instance, if $\widetilde{\sigma}_t^2$ is uniformly bounded, then (15) simply requires that $\widehat{\sigma}_t^2 - \widetilde{\sigma}_t^2 = o(t^{\eta-1})$ (i.e., strong consistency at *any* polynomial rate, potentially much slower than $t^{-1/2}$). Moreover, in the i.i.d. case with at least $(2+\delta)$ finite absolute moments, Condition L-3-η always holds by the SLLNs of Marcinkiewicz and Zygmund [25].

Our goal now is to show that sequences of AsympCSs given in Proposition 2.5 have asymptotic time-uniform coverage, and we will achieve this by effectively tuning them for later and later start times. Recall that Appendix B.2 allows us to choose the parameter $\rho > 0$ so that the AsympCS is tightest at some particular time—we will now choose ρ_m based on the first peeking time m as

$$\rho_m := \rho(\widehat{\sigma}_m^2 m \log(m \vee e)) \equiv \sqrt{\frac{-2\log\alpha + \log(-2\log\alpha) + 1}{\widehat{\sigma}_m^2 m \log(m \vee e)}}.[9]$$

Then, let $(\widetilde{C}_t(m))_{t=m}^{\infty}$ be the Gaussian mixture AsympCS with ρ_m plugged into the expression of the boundary for all $t \geq m$:

$$\widetilde{C}_t(m) := \left(\widehat{\mu}_t \pm \sqrt{\frac{t\widehat{\sigma}_t^2\rho_m^2 + 1}{t^2\rho_m^2}\log\left(\frac{t\widehat{\sigma}_t^2\rho_m^2 + 1}{\alpha^2}\right)}\right). \tag{16}$$

In other words, $(\widetilde{C}_t(m))_{t=m}^{\infty}$ should be thought of as an AsympCS that only starts after time m, and is vacuous beforehand. The following theorem formalizes the coverage guarantees satisfied by this *sequence* of AsympCSs as $m \to \infty$.

THEOREM 2.8 (Asymptotic $(1-\alpha)$-coverage for Gaussian mixture AsympCSs). *Given the same setup as Proposition 2.5 and Conditions L-1, L-2, and L-3-η, the AsympCSs $(\widetilde{C}_t(m))_{t=m}^{\infty}$ given in (16) have sharp asymptotic $(1-\alpha)$-coverage for $\widetilde{\mu}_t := \frac{1}{t}\sum_{i=1}^t \mu_i$ as $m \to \infty$, meaning*

$$\lim_{m\to\infty} \mathbb{P}(\forall t \geq m,\ \widetilde{\mu}_t \in \widetilde{C}_t(m)) = 1 - \alpha.$$

[9] In fact, ρ_m can be replaced by $\rho(\widehat{\sigma}_m^2 m d_m)$ where $(d_m)_{m=1}^{\infty}$ is any positive increasing sequence diverging to ∞.

The proof can be found in Appendix A.7. Clearly, when the mean is constant—i.e. $\mu_1 = \mu_2 = \cdots = \mu_\star$—the above also holds for the running intersection of intervals $\bigcap_{m \le s \le t} \widetilde{C}_s(m)$. Notice that in the i.i.d. setting, as $m \to \infty$, $(\widetilde{C}_t(m))_{t=m}^{\infty}$ is asymptotically equivalent to $(\overline{C}_t^{\mathcal{G}}(m))_{t=m}^{\infty}$ given by

$$\overline{C}_t^{\mathcal{G}}(m) := \left(\widehat{\mu}_t \pm \widehat{\sigma}_t \sqrt{\frac{t\overline{\rho}_m^2 + 1}{t^2 \overline{\rho}_m^2} \log\left(\frac{t\overline{\rho}_m^2 + 1}{\alpha^2}\right)}\right), \tag{17}$$

where $\overline{\rho}_m := \rho(m \log(m \vee e))$. A quick inspection of the proof will reveal that (17) also satisfies the coverage guarantee provided in Theorem 2.8 under the condition that $\widetilde{\sigma}_t^2 \to \sigma_\star^2 > 0$ almost surely. In summary, (17) can be thought of as an analogue of (16) for the AsympCSs that were derived in Theorem 2.2.

2.6. *Asymptotic confidence sequences using Robbins' delayed start.* As is clear from Theorem 2.4, virtually any boundary for Gaussian observations can be used to derive an AsympCS as long as an appropriate strong invariance principle can be applied under the given assumptions—indeed, Theorem 2.2, Proposition 2.3, Proposition 2.5, and Corollary 2.6 are all instantiations of the general phenomenon outlined in Theorem 2.4.

Another AsympCS that may be of interest to practitioners is one that leverages Robbins' CS for means of Gaussian random variables with a delayed start time [30], equation (20). In a nutshell, Robbins calculated a lower bound on the probability that a centered Gaussian random walk would remain within a particular two-sided boundary for all times $t \ge m$ given some starting time $m \ge 1$. That is, he showed that for i.i.d. Gaussians $(Z_t)_{t=1}^{\infty}$ with mean zero and variance σ^2, letting $G_t := \sum_{i=1}^{t} Z_i / \sigma$, and for any $a > 0$,

$$\mathbb{P}\left(\forall t \ge m : |G_t| < \sqrt{t\left(a^2 + \log(t/m)\right)}\right) \ge 1 - \Lambda(a), \tag{18}$$

where $\Lambda(a) := 2(1 - \Phi(a) + a\phi(a))$ and Φ and ϕ are the CDF and PDF of a standard Gaussian, respectively. In particular, setting $a \in (0, \infty)$ so that $\Lambda(a) = \alpha$ yields a two-sided $(1-\alpha)$-boundary for the Gaussian random walk $(G_t)_{t=1}^{\infty}$, and indeed, a solution to $\Lambda(a) = \alpha$ always exists and is trivial to compute due to the fact that Λ is strictly decreasing, starting at $\Lambda(0) = 1$ and $\lim_{a \to \infty} \Lambda(a) = 0$. In a followup paper, Robbins and Siegmund [31] extended the ideas of Robbins [30] to a large class of boundaries for Wiener processes so that the probabilistic inequality in (18) can be shown to be an equality when $|G_t|$ is replaced by the absolute value of a Wiener process (which would imply the inequality in (18) for i.i.d. standard Gaussians as a corollary). Using this fact within the general framework of Theorem 2.4 combined with the strong invariance principle of Strassen [40] and the techniques found in the proof of Theorem 2.8 yields the following result.

PROPOSITION 2.9 (Delayed-start AsympCS). *Consider the same setup as Theorem 2.8 so that $(Y_t)_{t=1}^{\infty}$ have conditional means and variances given by $\mu_t := \mathbb{E}(Y_t | Y_1^{t-1})$ and $\sigma_t^2 :=$ $\operatorname{var}(Y_t | Y_1^{t-1})$. Then under Conditions L-1, L-2, and L-3, we have that for any $m \ge 1$,*

$$\widetilde{C}_t^{\mathrm{DS}}(m) := \left(\widehat{\mu}_t \pm \widehat{\sigma}_t \sqrt{t^{-1} \cdot \left[a^2 + \log\left(t\widehat{\sigma}_t^2 / (m\widehat{\sigma}_m^2)\right)\right]}\right) \quad \text{if } t \ge m \tag{19}$$

(and all of $\mathbb{R}$ otherwise) forms a $(1-\alpha)$-AsympCS for $\widetilde{\mu}_t$, where a is chosen so that $\Lambda(a) = \alpha$. Furthermore, under Condition L-3-η, $\widetilde{C}_t^{\mathrm{DS}}(m)$ has sharp asymptotic time-uniform $(1-\alpha)$-coverage in the sense of Definition 2.7.

The proof is provided in Appendix A.9. Similar to the relationship between Theorem 2.8 and (17), we have that if variances happen to converge $\widetilde{\sigma}_t^2 \to \sigma_\star^2 > 0$ almost surely, then as a corollary of Proposition 2.9, the sequence $(\widetilde{C}_t^{\mathrm{DS}\star}(m))_{t=1}^{\infty}$ given by

$$\widetilde{C}_t^{\mathrm{DS}\star}(m) := \left(\widehat{\mu}_t \pm \widehat{\sigma}_t \sqrt{t^{-1} \cdot \left(a^2 + \log(t/m)\right)}\right) \quad \text{for } t \ge m \tag{20}$$

has asymptotic time-uniform coverage as in Definition 2.7. This can be seen as a generalization and improvement of results implied by Robbins [30] and Robbins and Siegmund [31], and has connections to Bibaut, Kallus and Lindon [2]. We elaborate on these below.

2.6.1. *Relationship to Robbins and Siegmund.* Informally, Robbins and Siegmund [31], Theorem 2(i), show that for independent and identically distributed random variables $(Y_t)_{t=1}^{\infty} \sim \mathbb{P}$ with mean zero and unit variance (without loss of generality), the probability of their scaled partial sums $S_t := \sum_{i=1}^{t} Y_i$ exceeding a particular boundary behaves like a rescaled Wiener process exceeding that boundary. Consequently, for i.i.d. data with known variance, their result combined with Robbins' delayed start [30], equation (20), implies an asymptotic coverage guarantee (in the sense of Definition 2.7) for $\widetilde{C}_t^{\mathrm{DS}\star}$ given in (20) but with $\widehat{\sigma}_t$ replaced by the true σ. As such, (20) should be thought of as a generalization of their boundary when variances are unknown under martingale dependence. Nevertheless, Proposition 2.9 is strictly more general, allowing for variances to never converge.

2.6.2. *Relationship to Bibaut*, *Kallus and Lindon* [2]. In version 1 of Bibaut, Kallus and Lindon [2], the authors derive a particular AsympCS and show that sequences thereof satisfy an asymptotic coverage guarantee in the same sense as Definition 2.7. That bound resembles—but is always looser than—a corollary of Proposition 2.9 in (20) for a fixed $m \geq 1$; see Claim B.1. Nevertheless, it was their asymptotic type-I error results that inspired us to show that the same guarantees hold for the bounds in (16), (19), and (20), and with our explicit conditions on how variances are allowed to diverge in the former two, rather than their implicit conditions (or sufficient conditions in special cases) through almost-surely convergent variance-stabilized pseudo-outcomes. Nevertheless, one can always apply our bounds to these same variance-stabilized pseudo-outcomes to weaken these implicit assumptions. After our advances, the second version of their paper introduced a bound called the "running maximum likelihood SPRT" (rmlSPRT) which is identical to the corollary of Proposition 2.9 found in (20) (modulo differences in variance estimation techniques); see Claim B.2. Since the exact connections may not be obvious to the reader, we derive them explicitly in Appendix B.7. Despite their rmlSPRT being identical to (20), we remark that their paper focuses on testing guarantees and contains several additional interesting investigations including a sophisticated analysis of the expected rejection time, enriching the landscape of asymptotic anytime-valid methodology.

3. Illustration: causal effects and semiparametric estimation.

Given the groundwork laid in Section 2, we now demonstrate the use of AsympCSs for conducting anytime-valid causal inference. Since it is an important and thoroughly studied functional, we place a particular emphasis on the average treatment effect (ATE) for illustrative purposes but we discuss how these techniques apply to semiparametric functional estimation more generally alongside a delta method for AsympCSs in Section 3.5. The literature on semiparametric functional inference often falls within the asymptotic regime and hence AsympCSs form a natural time-uniform extension thereof.

It is important to note that obtaining AsympCSs for the ATE is not as simple as directly applying the theorems of Section 2.1 to some appropriately chosen augmented inverse-probability-weighted (AIPW) influence functions (otherwise the illustration of this section would have been trivial). Indeed, satisfying the conditions of the aforementioned theorems—Theorem 2.4 in particular—in the presence of infinite-dimensional nuisance parameters is nontrivial and the analysis proceeds rather differently from the fixed-n setting. Nevertheless, after introducing and carefully analyzing sequential sample splitting and cross-fitting (Section 3.1), we will see that asymptotic time-uniform inference for the ATE is possible.

To solidify the notation and problem setup, suppose that we observe a (potentially infinite) sequence of i.i.d. variables $Z_1, Z_2, \dots$ from a distribution $\mathbb{P}$ where $Z_t := (X_t, A_t, Y_t)$ denotes the t^{th} subject's triplet and $X_t \in \mathbb{R}^d$ are their measured baseline covariates, $A_t \in \{0, 1\}$ is the treatment that they receive, and $Y_t \in \mathbb{R}$ is their measured outcome after treatment. Our target estimand is the average treatment effect (ATE) ψ defined as

$$\psi := \mathbb{E}(Y^1 - Y^0),$$

where Y^a is the counterfactual outcome for a randomly selected subject had they received treatment $a \in \{0, 1\}$. The ATE ψ can be interpreted as the average population outcome if everyone were treated $\mathbb{E}(Y^1)$ versus if no one were treated $\mathbb{E}(Y^0)$. Under standard causal identification assumptions—typically referred to as consistency, positivity, and exchangeability (see, e.g. Kennedy [18], §2.2)—we have that ψ can be written as a (non-counterfactual) functional of the distribution $\mathbb{P}$:

$$\psi \equiv \psi(\mathbb{P}) = \mathbb{E}\big\{\mathbb{E}(Y \mid X, A = 1) - \mathbb{E}(Y \mid X, A = 0)\big\}.$$

Throughout the remainder of this section, we will operate under these identification assumptions and aim to derive efficient AsympCSs for ψ using tools from semiparametric theory. At a high level, we will construct AsympCSs for ψ by combining the results of Section 2 with sample averages of *influence functions* for ψ and in the ideal case, these influence functions will be *efficient* (in the semiparametric sense).

3.1. *Sequential sample splitting and cross fitting.* Following Robins et al. [32], Zheng and van der Laan [52], and Chernozhukov et al. [6], we employ sample splitting (or cross fitting) to derive an estimate $\widehat{f}$ of the influence function f on a "training" sample, and evaluate $\widehat{f}$ on values of Z_t in an independent "evaluation" sample. Sample splitting sidesteps complications introduced from "double-dipping" (i.e., using Z_t to both construct $\widehat{f}$ and evaluate $\widehat{f}(Z_t)$) and simplifies the analysis of the downstream estimator. The aforementioned authors employed sample splitting in the *batch* (nonsequential) regime while we are concerned with settings where data are continually observed in an online stream over time, and hence we modify the sample splitting procedure as follows. We will denote $\mathcal{D}_\infty^{\text{trn}}$ and $\mathcal{D}_\infty^{\text{eval}}$ as the "training" and "evaluation" sets, respectively. At time t, we assign Z_t to either group with equal probability:

$$Z_t \in \begin{cases} \mathcal{D}_\infty^{\text{trn}} & \text{with probability } 1/2, \\ \mathcal{D}_\infty^{\text{eval}} & \text{otherwise.} \end{cases}$$

Note that at time $t + 1$, Z_t is *not* re-randomized into either split—once Z_t is randomly assigned to one of $\mathcal{D}_\infty^{\text{trn}}$ or $\mathcal{D}_\infty^{\text{eval}}$, they remain in that split for the remainder of the study. In this way, we can write $\mathcal{D}_\infty^{\text{trn}} = (Z_1^{\text{trn}}, Z_2^{\text{trn}}, \dots)$ and $\mathcal{D}_\infty^{\text{eval}} = (Z_1^{\text{eval}}, Z_2^{\text{eval}}, \dots)$ and think of these as independent, sequential observations from a common distribution $\mathbb{P}$. To keep track of how many subjects have been randomized to $\mathcal{D}_\infty^{\text{trn}}$ and $\mathcal{D}_\infty^{\text{eval}}$ at time t, define

$$T := \big|\mathcal{D}_\infty^{\text{eval}}\big| \quad \text{and} \quad T' := \big|\mathcal{D}_\infty^{\text{trn}}\big| \equiv t - T,$$

where we have left the dependence on t implicit.

REMARK 2. Strictly speaking, under the i.i.d. assumption, we do not need to randomly assign subjects to training and evaluation groups for the forthcoming results to hold (e.g. we could simply assign even-numbered subjects to $\mathcal{D}_\infty^{\text{trn}}$ and odd-numbered subjects to $\mathcal{D}_\infty^{\text{eval}}$). However, the analysis is not further complicated by this randomization, and it can be used to combat bias in treatment assignments when the i.i.d. assumption is violated [9].

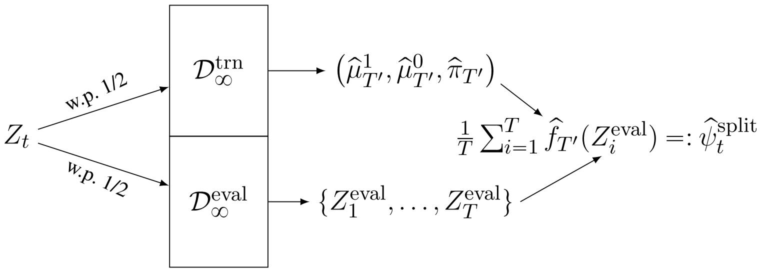
Z_t
w.p. 1/2
w.p. 1/2
$\mathcal{D}_\infty^{\text{trn}}$
$\mathcal{D}_\infty^{\text{eval}}$
$(\widehat{\mu}^1_{T'}, \widehat{\mu}^0_{T'}, \widehat{\pi}_{T'})$
$\{Z_1^{\text{eval}}, \dots, Z_T^{\text{eval}}\}$
$\frac{1}{T}\sum_{i=1}^T \widehat{f}_{T'}(Z_i^{\text{eval}}) =: \widehat{\psi}_t^{\text{split}}$

FIG. 3. *A schematic illustrating sequential sample splitting. At each time step t, the new observation Z_t is randomly assigned to $\mathcal{D}_\infty^{\text{trn}}$ or $\mathcal{D}_\infty^{\text{eval}}$ with equal probability* (1/2). *Nuisance function estimators $(\widehat{\mu}^1_{T'}, \widehat{\mu}^0_{T'}, \widehat{\pi}_{T'})$ are constructed using $\mathcal{D}_\infty^{\text{trn}}$ which then yield $\widehat{f}_{T'}$. The sample-split estimator $\widehat{\psi}_t^{\text{split}}$ is defined as the sample average $\frac{1}{T}\sum_{i=1}^T \widehat{f}_{T'}(Z_i^{\text{eval}})$ where each $Z_i^{\text{eval}} \in \mathcal{D}_\infty^{\text{eval}}$.*

The "sample split" estimator would then be given by $\widehat{\psi}_t^{\text{split}} := \frac{1}{T}\sum_{i=1}^T \widehat{f}_{T'}(Z_i^{\text{eval}})$ where $\widehat{f}_{T'}$ is an estimate of the so-called *efficient influence function* (a brief review of semiparametric efficient estimators can be found in Appendix B.9) given by

$$f(z) \equiv f(x,a,y) := \{\mu^1(x) - \mu^0(x)\} + \left(\frac{a}{\pi(x)} - \frac{1-a}{1-\pi(x)}\right)\{y - \mu^a(x)\}. \tag{21}$$

Crucially, $\widehat{f}_{T'}$ takes the form of (21) but with $\eta \equiv (\mu^1, \mu^0, \pi)$ replaced by $\widehat{\eta}_{T'} \equiv (\widehat{\mu}^1_{T'}, \widehat{\mu}^0_{T'}, \overline{\pi}_{T'})$—where $\overline{\pi}_{T'}$ may be an estimator $\widehat{\pi}_{T'}$ of the propensity score π, or the propensity score itself, depending on whether one is considering an observational study or randomized experiment—so that $\widehat{\eta}_{T'}$ is built solely from $\mathcal{D}_\infty^{\text{trn}}$. The sample splitting procedure for constructing $\widehat{\psi}_t^{\text{split}}$ is summarized pictorially in Figure 3. In the batch setting for a fixed sample size, $\widehat{\psi}_t^{\text{split}}$ is often referred to as the *augmented inverse probability weighted* (AIPW) estimator [34, 35] (an instantiation of so-called "one-step correction" in the semiparametrics literature) and we adopt similar nomenclature here. However, a commonly cited downside of sample splitting is the loss in efficiency by using $T \approx t/2$ subjects instead of t when evaluating the sample mean $\frac{1}{T}\sum_{i=1}^T \widehat{f}_{T'}(Z_i^{\text{eval}})$. An easy fix is to *cross-fit*: swap the two samples, using $\mathcal{D}_\infty^{\text{eval}}$ for training and $\mathcal{D}_\infty^{\text{trn}}$ for evaluation to recover the full sample size of $t \equiv T + T'$ [6, 32, 52]. That is, construct $\widehat{f}_T$ solely from $\mathcal{D}_\infty^{\text{eval}}$ and define the cross-fit estimator $\widehat{\psi}_t^\times$ as

$$\widehat{\psi}_t^\times := \frac{\sum_{i=1}^T \widehat{f}_{T'}(Z_i^{\text{eval}}) + \sum_{i=1}^{T'} \widehat{f}_T(Z_i^{\text{trn}})}{t}, \tag{22}$$

and the associated cross-fit variance estimate

$$\widehat{\operatorname{var}}_t(\widehat{f}) := \frac{\widehat{\operatorname{var}}_T(\widehat{f}_{T'}) + \widehat{\operatorname{var}}_{T'}(\widehat{f}_T)}{2}, \tag{23}$$

where $\widehat{\operatorname{var}}_T(\widehat{f}_{T'})$ is the $\mathcal{D}_\infty^{\text{eval}}$-sample variance of the pseudo-outcomes $(\widehat{f}_{T'}(Z_i^{\text{eval}}))_{i=1}^T$ and similarly for $\widehat{\operatorname{var}}_{T'}$ (we deliberately omit the subscript on $\widehat{f}$ in the left-hand side of (23)). For simplicity, all of the results that follow are stated in terms of the cross-fit estimators $(\widehat{\psi}_t^\times)_{t=1}^\infty$. With the setup of Appendix B.9 and Section 3.1 in mind, we are ready to derive AsympCSs for ψ, first in randomized experiments.

3.2. *Asymptotic confidence sequences in randomized experiments.* Consider a sequential randomized experiment so that a subject with covariates X has a known propensity score

$$\pi(X) := \mathbb{P}(A = 1 \mid X).$$

Consider the cross-fit AIPW estimator $\widehat{\psi}_t^{\times}$ as given in (22) but with estimated propensity scores—$\widehat{\pi}_{T'}(x)$ and $\widehat{\pi}_T(x)$—replaced by their true values $\pi(x)$, and with $\widehat{\mu}_{T'}^a$ and $\widehat{\mu}_T^a$ being possibly misspecified estimators for μ^a. We will assume that $\widehat{\mu}_t^a$ converges to some function $\overline{\mu}^a$, which need not coincide with μ^a. In what follows, when we use $\widehat{\mu}_t^a$ or $\widehat{f}_t$ in writing $\|\widehat{\mu}_t^a - \overline{\mu}^a\|_{L_2(\mathbb{P})}$ or $\|\widehat{f}_t - \overline{f}\|_{L_2(\mathbb{P})}$, we are referring to large-sample properties of the estimator (and hence $\widehat{f}_t$ could be replaced by $\widehat{f}_{T'}$ or $\widehat{f}_T$ without loss of generality).

THEOREM 3.1 (AsympCSs for the ATE in randomized experiments). *Let $\widehat{\psi}_t^{\times}$ be the cross-fit AIPW estimator as in* (22). *Suppose $\|\widehat{\mu}_t^a(X) - \overline{\mu}^a(X)\|_{L_2(\mathbb{P})} = o(1)$ for each $a \in \{0, 1\}$ where $\overline{\mu}^a$ is some function* (*but need not be μ^a*), *and hence $\|\widehat{f}_t - \overline{f}\|_{L_2(\mathbb{P})} = o(1)$ for some influence function $\overline{f}$. Suppose that propensity scores are bounded away from* 0 *and* 1, *i.e.* $\pi(X) \in [\delta, 1-\delta]$ *almost surely for some* $\delta > 0$, *and suppose that* $\text{var}(\overline{f}(Z)) < \infty$. *Then for any constant* $\rho > 0$,

$$\widehat{\psi}_t^{\times} \pm \sqrt{\widehat{\text{var}}_t(\widehat{f})} \cdot \sqrt{\frac{t\rho^2 + 1}{t^2\rho^2} \log\left(\frac{t\rho^2+1}{\alpha^2}\right)} \tag{24}$$

forms a $(1-\alpha)$*-AsympCS for* ψ.

The proof in Appendix A.5 combines an analysis of the almost-sure convergence of $(\widehat{\psi}_t^{\times} - \psi)$ with the AsympCS of Theorem 2.2. Notice that since $\widehat{\mu}_t^a$ is consistent for a function $\overline{\mu}^a$, we have that $\widehat{f}_t$ is converging to some influence function $\overline{f}$ of the form

$$\overline{f}(z) \equiv \overline{f}(x, a, y) := \{\overline{\mu}^1(x) - \overline{\mu}^0(x)\} + \left(\frac{a}{\pi(x)} - \frac{1-a}{1-\pi(x)}\right)\{y - \overline{\mu}^a(x)\}.$$

In practice, however, one must choose $\widehat{\mu}_t^a$. As alluded to at the beginning of Section 3, the best possible influence function is the EIF $f(z)$ defined in (21), and thus it is natural to attempt to construct $\widehat{\mu}_t^a$ so that $\|\widehat{f}_t - f\|_{L_2(\mathbb{P})} = o(1)$. The resulting AsympCSs would inherit such optimality properties, a point which we discuss further in Appendix B.10.

Since μ^a is simply a conditional mean function, we can use virtually any regression techniques to estimate it. Here we will consider the general approach of *stacking* introduced by Breiman [4] and further studied by Tsybakov [42] and van der Laan, Polley and Hubbard [44] (see also [29]) under the names of "aggregation" and "Super Learning" respectively. In short, stacking uses cross-validation to choose a weighted combination of K candidate predictors where the weights are chosen based on data in held-out samples. Importantly (and under certain conditions), the stacked predictor will have a mean squared error that scales with that of the best of the K predictors up to an additive $\log K$ term [42, 46]. This advantage can be seen empirically in Figure 4 where the true regression functions μ^0 and μ^1 are nonsmooth and nonlinear in x. Such advantages via stacking are not new—we are only highlighting the observation that similar phenomena carry over to AsympCSs.

So far, the use of flexible regression techniques like stacking were used only for the purposes of deriving sharper AsympCSs in sequential randomized experiments. In observational studies, however, consistent estimation of nuisance functions at fast rates is essential to the construction of *valid* fixed-n CIs, and indeed the same is true for AsympCSs.

3.3. *Asymptotic confidence sequences in observational studies.* Consider now a sequential observational study (e.g., we are able to continuously monitor $(X_t, A_t, Y_t)_{t=1}^{\infty}$ but do not know $\pi(x)$ exactly, or we are in a sequentially randomized experiment with noncompliance, etc.). The only difference in this setting with respect to setup is the fact that $\pi(x)$ is no longer known and must be estimated. As in the fixed-n setting, this complicates estimation and inference. The following theorem provides the conditions under which we can construct AsympCSs for ψ using the cross-fit AIPW estimator in observational studies.

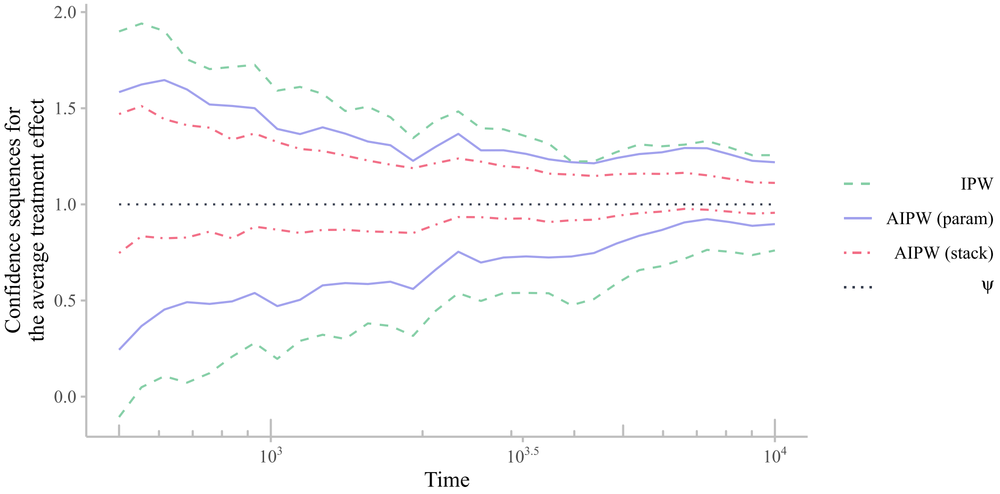

FIG. 4. *Three 90%-AsympCSs for the average treatment effect in a simulated randomized experiment using different regression estimators. Notice that all three AsympCSs uniformly capture the average treatment effect ψ, but more sophisticated models do so more efficiently, with AIPW+stacking greatly outperforming IPW.*

THEOREM 3.2 (AsympCSs for the ATE in observational studies). *Consider the same setup as Theorem 3.1 but with $\pi(x)$ unknown. Suppose that regression functions and propensity scores are consistently estimated in $L_2(\mathbb{P})$ at a product rate of $o(\sqrt{\log t/t})$, meaning*

$$\|\widehat{\pi}_t - \pi\|_{L_2(\mathbb{P})} \sum_{a=0}^{1} \|\widehat{\mu}_t^a - \mu^a\|_{L_2(\mathbb{P})} = o(\sqrt{\log t/t}).$$

Moreover, suppose that $\|\widehat{f}_t - f\|_{L_2(\mathbb{P})} = o(1)$ where f is the efficient influence function (21) *and that* $\operatorname{var}(f(Z)) < \infty$. *Then for any constant* $\rho > 0$, (24) *forms a* $(1-\alpha)$*-AsympCS for* ψ.

The proof in Appendix A.5.2 proceeds similarly to the proof of Theorem 3.1 by combining Theorem 2.2 with an analysis of the almost-sure behavior of $(\widehat{\psi}_t^\times - \psi)$. Notice that the nuisance estimation rate of $\sqrt{\log t/t}$ is slower than $1/\sqrt{t}$ which is usually required in the fixed-n regime, but we do require almost-sure convergence rather than convergence in probability.

Unlike the experimental setting of Section 3.2, Theorem 3.2 requires that $\widehat{\mu}_t^a$ and $\widehat{\pi}_t$ consistently estimate μ^a and π, respectively. As a consequence, the stacking-based AIPW AsympCS is both the tightest of the three *and* is uniquely consistent for ψ (see Figure 5).

3.4. *The running average of individual treatment effects.* The results in Sections 3.2 and 3.3 considered the classical regime where the ATE ψ is a fixed functional that does not change over time. Consider a strict generalization where distributions—and hence individual treatment effects in particular—may change over time. In other words,

$$\psi_t := \mathbb{E}\{Y_t^1 - Y_t^0\} \overset{(\star)}{=} \mathbb{E}\{\mathbb{E}(Y_t | X_t, A_t = 1) - \mathbb{E}(Y_t | X_t, A_t = 0)\},$$

where the equality $(\star)$ holds under the familiar causal identification assumptions discussed earlier. Despite the nonstationary and non-i.i.d. structure, it is nevertheless possible to derive AsympCSs for the *running average* of individual treatment effects $\widetilde{\psi}_t := \frac{1}{t}\sum_{i=1}^t \psi_i$—or simply, the running average treatment effect—using the Lyapunov-type bounds of Corollary 2.6. However, given this more general setup, the assumptions required are more subtle (but no more restrictive) than those for Theorems 3.1 and 3.2; as such, we describe their details here but handle the randomized and observational settings simultaneously for brevity.

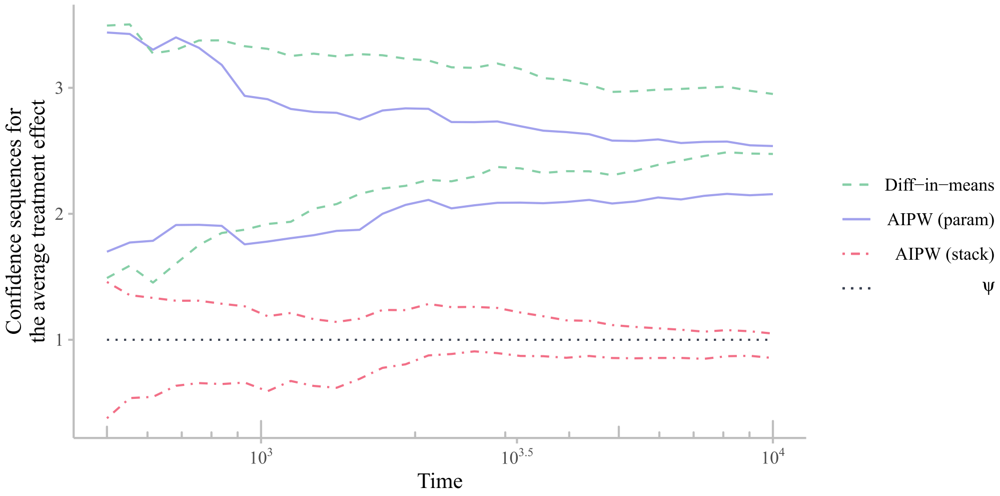

FIG. 5. *Three 90%-AsympCSs for the ATE in an observational study using three different estimators—a difference-in-means estimator, AIPW with parametric models, and AIPW with an ensemble of predictors combined via stacking. Unlike the randomized setup, only the stacking ensemble is consistent, since the other two are misspecified. Not only is the stacking-based AsympCS converging to ψ, but it is also the tightest of the three models at each time step.*

CONDITION $\widetilde{\text{ATE}}$-1 (Regression estimator is uniformly well-behaved in $L_2(\mathbb{P})$). *We assume that regression estimators $\mu_t^a(X_i)$ converge in $L_2(\mathbb{P})$ to any function $\overline{\mu}^a(X_i)$ uniformly for $i \in \{1, 2, \dots\}$, that is,*

$$\sup_{1 \le i < \infty} \|\widehat{\mu}_t^a(X_i) - \overline{\mu}^a(X_i)\|_{L_2(\mathbb{P})} = o(1) \quad \textit{for each } a \in \{0, 1\}.$$

Condition $\widetilde{\text{ATE}}$-1 simply requires that the regression estimator $\widehat{\mu}_t^a$ must converge to some function $\overline{\mu}^a$, which need not coincide with true regression function μ^a. In the i.i.d. setting where $X_1, X_2, \dots$ all have the same distribution, we would simply drop the $\sup_{1 \le i \le \infty}$, recovering the conditions for Theorems 3.1 and 3.2.

CONDITION $\widetilde{\text{ATE}}$-2 (Convergence of average nuisance errors). *Let $\widehat{\mu}_t^a$ be an estimator of the regression function μ^a, $a \in \{0, 1\}$ and $\widehat{\pi}_t$ an estimator of the propensity score π. We assume that the average bias shrinks at a $\sqrt{\log t / t}$ rate, that is,*

$$\frac{1}{t} \sum_{i=1}^{t} \left\{ \|\widehat{\pi}_t(X_i) - \pi(X_i)\|_{L_2(\mathbb{P})} \sum_{a=0}^{1} \|\widehat{\mu}_t^a(X_i) - \mu^a(X_i)\|_{L_2(\mathbb{P})} \right\} = o\left(\sqrt{\frac{\log t}{t}} \right). \tag{25}$$

Note that Condition $\widetilde{\text{ATE}}$-2 would hold in two familiar scenarios. First, in a randomized experiment (Theorem 3.3) where $\widehat{\pi}_t = \pi$ is known by design, we have that (25) is always zero, satisfying Condition $\widetilde{\text{ATE}}$-2 trivially. Second, in an observational study where the product of errors $\|\widehat{\pi}_t(X_i) - \pi(X_i)\|_{L_2(\mathbb{P})} \|\widehat{\mu}_t^a(X_i) - \mu^a(X_i)\|_{L_2(\mathbb{P})}$ vanishes at a rate faster than $\sqrt{\log t / t}$, for each i and for both $a \in \{0, 1\}$, we also have that their average product errors vanish at the same rate (25). With these assumptions in mind, let us summarize how running average treatment effects can be captured in randomized experiments.

THEOREM 3.3 (AsympCSs for the running average treatment effect). *Suppose $Z_1, Z_2, \dots$ are independent triples $Z_t := (X_t, A_t, Y_t)$ and that Conditions $\widetilde{\text{ATE}}$-1 and $\widetilde{\text{ATE}}$-2 hold.*

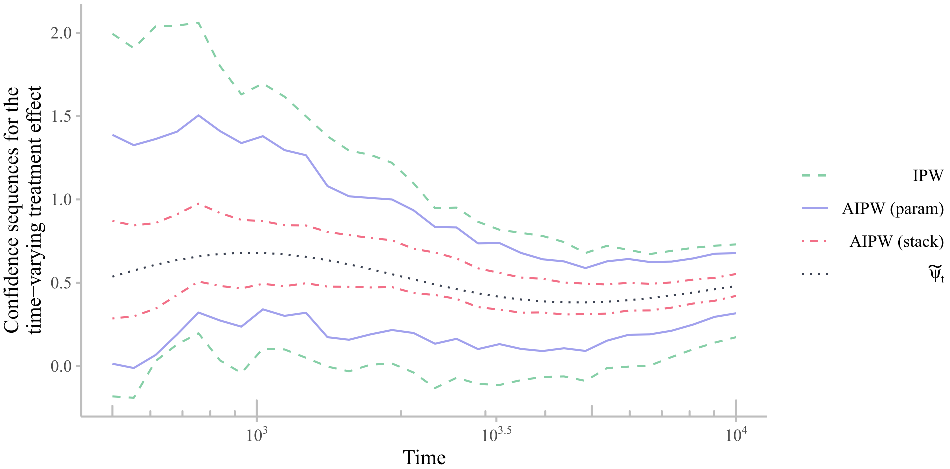

FIG. 6. *Three 90% AsympCSs for $\widetilde{\psi}_t$ constructed using various estimators via Theorem 3.3. Since this is a randomized experiment, all three CSs capture $\widetilde{\psi}_t$ uniformly over time with high probability. Similar to Figure 4, however, the stacking-based AIPW estimator greatly outperforms those based on parametric models or IPW.*

Finally, suppose that the conditions of Corollary 2.6 hold, but with $(Y_t)_{t=1}^{\infty}$ replaced by the influence functions $(\overline{f}(Z_t))_{t=1}^{\infty}$. Then

$$\widehat{\psi}_t^{\times} \pm \sqrt{\frac{t\rho^2\widehat{\operatorname{var}}_t(\overline{f})+1}{t^2\rho^2}\log\left(\frac{t\rho^2\widehat{\operatorname{var}}_t(\overline{f})+1}{\alpha^2}\right)} \tag{26}$$

forms a $(1-\alpha)$*-AsympCS for the running average treatment effect* $\widetilde{\psi}_t := \frac{1}{t}\sum_{i=1}^{t}\psi_i$.

The proof can be found in Appendix A.6. Note that both Theorems 3.1 and 3.2 are particular instantiations of Theorem 3.3. The important takeaway from Theorem 3.3 is that under some rather mild conditions on the moments of $(\overline{f}(Z_t))_{t=1}^{\infty}$, it is possible to derive an AsympCS for a running average treatment effect $\widetilde{\psi}_t$ (see Figure 6 for what these look like in practice). Nevertheless, under the commonly considered regime where the treatment effect is constant $\psi_1 = \psi_2 = \cdots = \psi$, we have that (26) forms a $(1-\alpha)$-AsympCS for ψ. Note that unlike Theorems 3.1 and 3.2, Theorem 3.3 actually does require the use of the cross-fit AIPW estimator $\widehat{\psi}_t^{\times}$ and would not capture $\widetilde{\psi}_t$ if the sample-split version were used in its place.

REMARK 3 (Avoiding sample splitting via martingale AsympCSs). The reader may wonder whether it is possible to simply plug in a *predictable* estimate of $\widehat{\mu}_t^a$ in a randomized experiment—that is, so that $\widehat{\mu}_t^a$ only depends on Z_1^{t-1}—and employ the Lindeberg-type martingale AsympCS of Proposition 2.5 in place of Corollary 2.6, thereby sidestepping the need for sequential sample splitting and cross fitting altogether. Indeed, such an analogue of Theorem 3.3 is possible to derive, but the conditions required are less transparent than those we have provided above so we defer it to Appendix B.6.

3.5. *Extensions to general semiparametric estimation and the delta method.* The discussion thus far has been focused on deriving AsympCSs for the ATE. However, the tools presented in this paper are more generally applicable to any pathwise differentiable functional with positive and finite semiparametric information bound. Some prominent examples in causal inference include modified interventions, complier-average effects, time-varying effects, and controlled mediation effects, among others. Examples outside causal inference include the expected density, entropy, the expected conditional variance, and the expected conditional covariance, to list a few. All of the aforementioned problems, including estimation of the ATE can be written in the following general form. Suppose $Z_1, Z_2, \ldots \sim \mathbb{Q}$ and let $\theta(\mathbb{Q})$ be some functional (such as those listed above) of the distribution $\mathbb{Q}$. In the case of a finite sample size n, $\widehat{\theta}_n$ is said to be an asymptotically linear estimator [41] for θ if the centered estimator $\widehat{\theta}_n - \theta$ can be written as a sample average of influence functions ϕ up to something vanishing in $\mathbb{Q}$-probability at a rate faster than $1/\sqrt{n}$, or in other words $\widehat{\theta}_n - \theta = \frac{1}{n}\sum_{i=1}^{n} \phi(Z_i) + o_{\mathbb{Q}}(1/\sqrt{n})$ When the sample size is not fixed in advance, we may analogously say that $\widehat{\theta}_t$ is an *asymptotically linear time-uniform estimator* if instead,

$$\widehat{\theta}_t - \theta = \frac{1}{t}\sum_{i=1}^{t} \phi(Z_i) + o(\sqrt{\log t/t}) \tag{27}$$

$\mathbb{Q}$-almost surely, with ϕ being the same influence function as before. For example, in the case of the ATE with $(Z_t)_{t=1}^{\infty} \sim \mathbb{P}$, we presented an efficient estimator $\widehat{\psi}_t$ for ψ which took the form of (27) with $\widehat{\theta}_t = \widehat{\psi}_t$, $\theta = \psi$, and $\phi(z) = f(z) - \psi$ where f is the uncentered efficient influence function (EIF) given in (21). In order to justify that the remainder term is indeed $o(\sqrt{\log t/t})$, we used sequential sample splitting and additional analysis in the randomized and observational settings. In general, as long as an estimator $\widehat{\theta}_t$ for θ has the form (27), we may derive AsympCSs for θ as a simple corollary of Theorem 2.2 with $\widehat{\mu}_t$ replaced by $\widehat{\theta}_t$ and $\widehat{\sigma}_t$ replaced by $\operatorname{var}(\phi)$. If $\widehat{\theta}_t$ involves nuisance parameters such as in Theorems 3.1 and 3.2, this can be handled on a case-by-case basis where sequential sample splitting and cross fitting (Section 3.1) may be helpful. We now derive an analogue of the delta method for asymptotically linear time-uniform estimators.

PROPOSITION 3.4 (The delta method for AsympCSs). *Let $\widehat{\theta}_t$ be an asymptotically linear time-uniform estimator of θ with influence function ϕ and let $g : \mathbb{R} \to \mathbb{R}$ be a continuously differentiable function with first derivative g'. Then, $g(\widehat{\theta}_t)$ is an asymptotically linear time-uniform estimator for $g(\theta)$ with influence function given by $g'(\theta)\phi(\cdot)$, i.e.*

$$g(\widehat{\theta}_t) - g(\theta) = \frac{1}{t}\sum_{i=1}^{t} g'(\theta)\phi(Z_i) + o(\sqrt{\log t/t}).$$

The short proof in Appendix A.8 is similar to the proof of the classical delta method but with the almost-sure continuous mapping theorem used in place of the in-probability one, and with the law of the iterated logarithm used in place of the central limit theorem.

4. Simulation studies: Widths and empirical coverage.

We now discuss the results of some simulations, the first of which illustrates the sharpness benefits of AsympCSs over nonasymptotic CSs in randomized experiments (Figure 7), while the second displays the empirical miscoverage rates for mean estimation as the tuning parameter ρ_m from Section 2.5 is tuned for larger and larger initial peeking (or "burn-in") times (Figure 8).

Following the motivations of Section 3, consider the problem of average treatment effect estimation in an experiment with $\{0, 1\}$-valued outcomes where all subjects are randomly assigned to treatment or control with equal probability 1/2. Since all propensity scores are given by 1/2, the estimated influence functions appearing in and surrounding (21) lie in $[-2, 2]$ almost surely, and thus the techniques of Robbins [30] and Howard et al. [14] for mean estimation can be used to construct nonasymptotic CSs for the ATE (as outlined by Howard et al. [14], §4.2). Figure 7 displays confidence sets and cumulative miscoverage rates for ATE estimation using (a) our AsympCSs, (b) CLT-based CIs, (c) the sub-exponential CSs of Howard et al. [14], §4.2, and (d) the sub-Gaussian CSs of Robbins [30]. Notice that the CLT-based CIs have cumulative miscoverage rates that quickly diverge beyond $\alpha = 0.1$ while those of

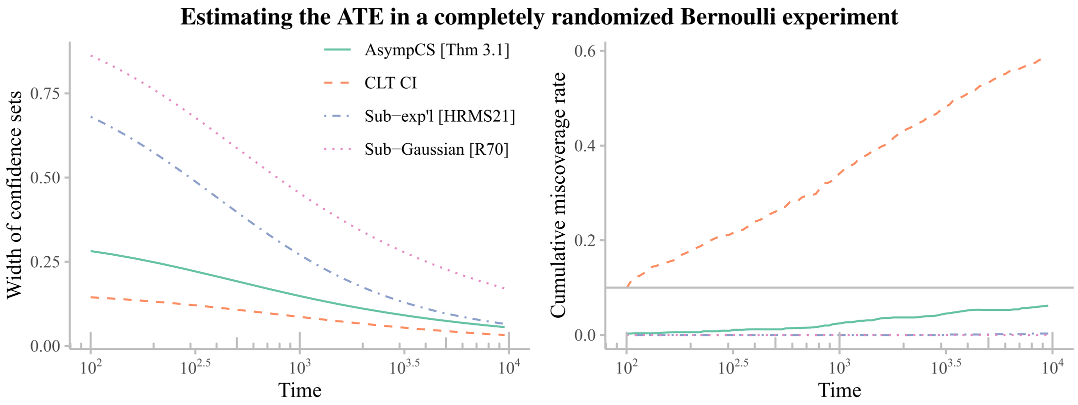

FIG. 7. *A comparison of* $(1-\alpha) \equiv 90\%$ *confidence sets for the ATE in a completely randomized Bernoulli experiment. Empirical widths and miscoverage rates were computed with* 1000 *replications beginning at time* 500. *Notice that only* (*asymptotic and nonasymptotic*) *CSs have miscoverage rates below* α, *but AsympCSs are the only ones that appear to sharply approach this level. The tuning parameter* $\overline{\rho}_m$ *was chosen for a start time of* m *as* $\overline{\rho}_m := \rho(m \log(m \vee e))$ *following* (17).

CSs—both asymptotic and nonasymptotic—never exceed α before time 10^4.[10] Moreover, notice that nonasymptotic CSs appear to be conservative, while our AsympCSs are much tighter and have miscoverage rates approaching α (as expected in light of Theorem 2.8).

Indeed, Figure 8 empirically and visually illustrates Theorem 2.8 by displaying the cumulative miscoverage rate of (16) for estimating the mean of Uniform(0, 1) and t-distributed random variables when tuned for later and later initial peeking times $m \in \{2, 5, 10, 50, 100\}$. Notice that as m increases, the cumulative miscoverage rate is uniformly contained below $\alpha = 0.1$ for all $m \leq t \leq 10^5$.

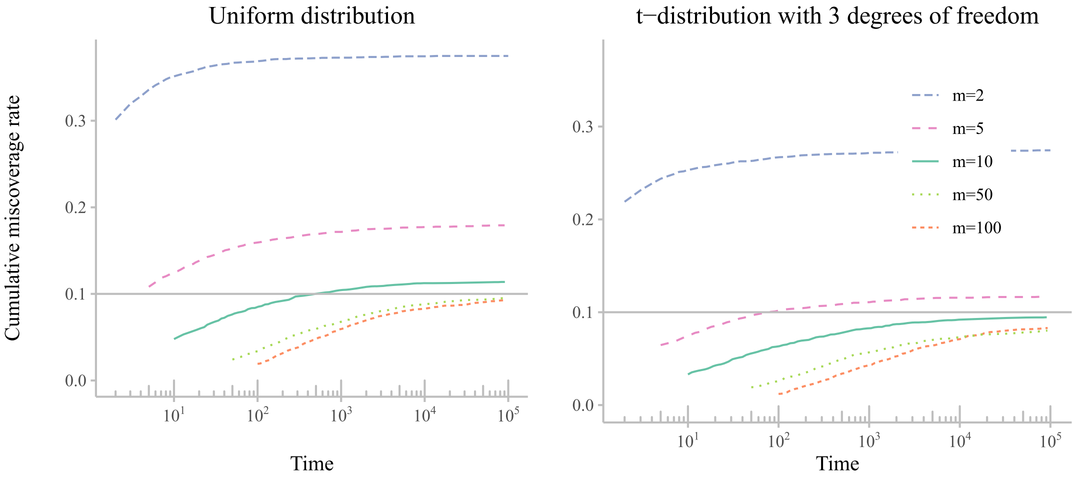

FIG. 8. *Cumulative miscoverage rates using* (17) *at level* $\alpha = 0.1$ *to estimate the mean of i.i.d. Uniform*(0, 1) *and* t*-distributed random variables in the left-hand and right-hand side plots*, *respectively. Notice that in both cases including the heavy-tailed setting of a* t*-distribution with* 3 *degrees of freedom* (*so that the variance is finite but third and higher absolute moments are all infinite*), *cumulative miscoverage rates do not exceed* $\alpha = 0.1$ *even after* 10^5 *observations as long as the first peeking time* m *is at least* 50. *It is worth remarking that asymptotic approximations appear to "kick in" earlier for the heavy-tailed* t*-distribution.*

[10] Note that a longer time horizon of 10^5 is considered only for AsympCSs and CLT-based CIs in Figure 1, but the shorter horizon of 10^4 is used here due to the computational expense of Howard et al. [14], Thm 2, at large t.

The setting considered in Figure 7 is one where AsympCSs have substantial benefits over nonasymptotic CSs because the latter suffer due to slightly conservative almost-sure bounds of $[-2, 2]$ on the (estimated) influence functions, while AsympCSs can adapt to the true variance irrespective of said *a priori* known bounds. However, the setting where all propensity scores are given by 1/2 is in some sense the "easiest" for nonasymptotic CSs, and the benefits of AsympCSs are exaggerated when those almost-sure influence function bounds increase relative to the true variance. Indeed, Appendix B.8 contains a more comprehensive set of simulations, including the simpler problem of univariate mean estimation for bounded random variables as well as treatment effect estimation in randomized experiments with "personalized" randomization, that is, covariate-dependent propensity scores with extreme almost-sure bounds but small variances; see Figure 2 for details.

Finally, while Figure 7 illustrates the benefits of AsympCSs over nonasymptotic CSs with respect to *tightness*, we also wish to highlight their benefits of *versatility*. In particular, there are many settings for which no simulations could have been run since AsympCSs provide the first (asymptotically) time-uniform solution in the literature. For example, as a consequence of Bahadur and Savage [1], it is impossible to derive nonasymptotic CSs (or CIs) for the mean of random variables without *a priori* known bounds on their moments. By contrast, AsympCSs can (much like CLT-based CIs) handle mean estimation under finite (but unknown) moment assumptions. It is also impossible to derive nonasymptotic CS (and CIs) for the ATE from observational studies (without unrealistic knowledge of nuisance function estimation errors) but Section 3.3 outlines an asymptotically time-uniform solution. In both settings, we do not run simulations akin to Figure 7 since there do not exist prior CSs to compare to.

5. Real data application: Effects of IV fluid caps in sepsis patients.

Let us now illustrate the use of Theorem 3.2 by sequentially estimating the effect of fluid-restrictive strategies on mortality in an observational study of real sepsis patients. We will use data from the Medical Information Mart for Intensive Care III (MIMIC-III), a freely available database consisting of health records associated with more than 45,000 critical care patients at the Beth Israel Deaconess Medical Center [17, 28]. The data contain demographics, vital signs, medications, and mortality, among other information collected over the span of 11 years.

Following Shahn et al. [36], we aim to estimate the effect of restricting intravenous (IV) fluids within 24 hours of intensive care unit (ICU) admission on 30-day mortality in sepsis patients. In particular, we considered patients at least 16 years of age satisfying the Sepsis-3 definition—that is, those with a suspected infection and a Sequential Organ Failure Assessment (SOFA) score of at least 2 [37]. Sepsis-3 patients can be obtained from MIMIC-III using SQL scripts provided by Johnson and Pollard [16], but we provide detailed instructions for reproducing our data collection and analysis process in our Supplementary Material [49]. This resulted in a total of 5231 sepsis patients, each of whom received out-of-hospital followup of at least 90 days.

Consider IV fluid intake within 24 hours of ICU admission $\mathscr{L}^{24\text{h}}$. To construct a binary treatment $A \in \{0, 1\}$, we dichotomize $\mathscr{L}^{24\text{h}}$ so that $A_i = \mathbb{1}(\mathscr{L}_i^{24\text{h}} \leq 6\text{L})$. 30-day mortality Y is defined as 1 if the patient died within 30 days of hospital admission, and 0 otherwise. We will consider baseline covariates X including a patient's age and sex, whether they are diabetic, modified Elixhauser scores [43], and SOFA scores. We consider the causal estimand

$$\psi := \mathbb{P}(Y^{\mathscr{L}^{24\text{h}} \leq 6L} = 1) - \mathbb{P}(Y^{\mathscr{L}^{24\text{h}} > 6L} = 1),$$

that is, the difference in average 30-day mortality that would be observed if all sepsis patients were randomly assigned an IV fluid level according to the lower truncated distribution $\mathbb{P}(\mathscr{L}^{24\text{h}} \leq l | \mathscr{L}^{24\text{h}} \leq 6\text{L})$ versus the upper truncated distribution $\mathbb{P}(\mathscr{L}^{24\text{h}} \leq l | \mathscr{L}^{24\text{h}} > 6\text{L})$

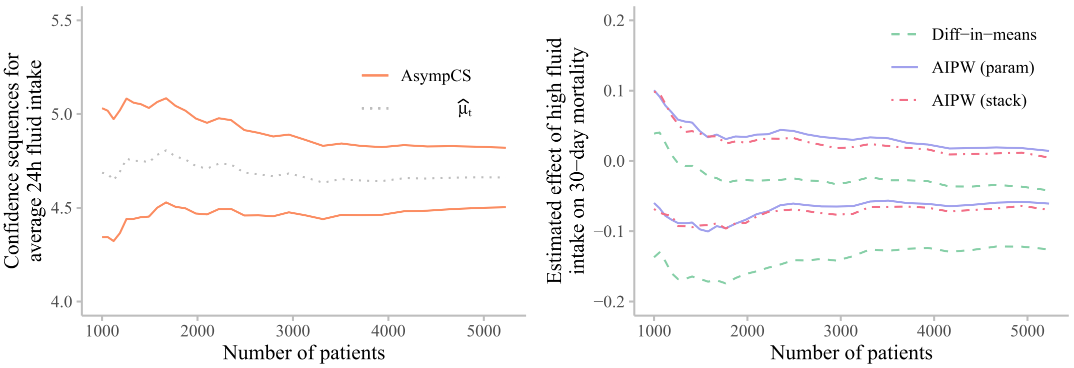

FIG. 9. *Left-hand side: a* 90% *AsympCS used to track the average* 24 *h fluid intake over time. Right-hand side: Three* 90%*-AsympCSs for the causal effect of capped IV fluid intake* (*defined as* $\leq$ 6 *litres*) *on* 30*-day mortality using the same three estimators as those outlined in Figure* 5. *Notice that an analysis using a difference-in-means estimator would conclude that the treatment effect is negative after observing fewer than* 1500 *patients.*

[8]. While this is technically a stochastic intervention effect, we have that under the same causal identification assumptions discussed in Section 3, ψ is identified as

$$\psi = \mathbb{E}\{\mathbb{E}(Y \mid X, A = 1) - \mathbb{E}(Y \mid X, A = 0)\},$$

the same functional considered in the previous sections. Therefore, we can estimate ψ under the same assumptions and with the same techniques as Section 3.3. Figure 9 contains AsympCSs for ψ using difference-in-means, parametric AIPW, and stacking-based AIPW estimators to demonstrate the impacts of different modeling choices on AsympCS width. Note that these simple binary treatment and outcome variables were used for simplicity so that the methods outlined in Section 3.3 are immediately applicable, but Section 3.5 points out that our AsympCSs may be used to sequentially estimate other causal functionals.

The stacking-based AIPW AsympCSs cover the null treatment effect of 0 from the 1000th to the 5231st observed patient, and thus we cannot conclude whether 6 L IV fluid caps have an effect on 30-day mortality in sepsis patients.

Note that these stacking-based AsympCSs nearly drop below 0 after observing the 5231st patient's outcome. If we were using fixed-time confidence intervals, the analyst would need to resist the temptation to resume data collection (e.g., to see whether the null $H_0 : \psi = 0$ could be rejected with a larger sample size) as this would inflate type-I error rates as seen in Figure 1. On the other hand, AsympCSs permit precisely this form of continued sampling.

6. Conclusion. This paper introduced the notion of an "asymptotic confidence sequence" as the time-uniform analogue of an asymptotic confidence interval based on the central limit theorem. We derived an explicit universal asymptotic confidence sequence for the mean from i.i.d. observations under weak moment assumptions by appealing to strong invariance principles. These results were extended to the setting where observations' distributions (including means and variances) can vary over time under martingale dependence, such that our asymptotic confidence sequences capture a moving parameter—the running average of the conditional means so far. We then applied the aforementioned results to the problem of doubly robust sequential inference for the average treatment effect in both randomized experiments and observational studies under i.i.d. sampling. Finally, we showed how these causal applications remain valid in the non-i.i.d. setting where distributions change over time, in which case our asymptotic confidence sequences capture a running average of individual treatment effects. The aforementioned results will enable researchers to continuously monitor sequential experiments—such as clinical trials and online A/B tests—as well as sequential observational studies even if treatment effects do not remain stationary over time.

SUPPLEMENTARY MATERIAL

Supplement to "Time-uniform central limit theory and asymptotic confidence sequences" [49] (DOI: 10.1214/24-AOS2408SUPPA; .pdf). Contains proofs of the main results, additional discussions, and figures.

Supplement to "Time-uniform central limit theory and asymptotic confidence sequences" (DOI: 10.1214/24-AOS2408SUPPB; .zip). An R package alongside instructions and code to reproduce data and figures used in the main paper.

REFERENCES

[1] Bahadur, R. R. and Savage, L. J. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Stat.* **27** 1115–1122. MR0084241 https://doi.org/10.1214/aoms/1177728077

[2] Bibaut, A., Kallus, N. and Lindon, M. (2022). Near-optimal non-parametric sequential tests and confidence sequences with possibly dependent observations. arXiv preprint. Available at arXiv:2212.14411.

[3] Billingsley, P. (1995). *Probability and Measure*, 3rd ed. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. A Wiley-Interscience Publication. MR1324786

[4] Breiman, L. (1996). Stacked regressions. *Mach. Learn.* **24** 49–64.

[5] Chatterjee, S. (2012). A new approach to strong embeddings. *Probab. Theory Related Fields* **152** 231–264. MR2875758 https://doi.org/10.1007/s00440-010-0321-8

[6] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

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

[8] Díaz Muñoz, I. and van der Laan, M. (2012). Population intervention causal effects based on stochastic interventions. *Biometrics* **68** 541–549. MR2959621 https://doi.org/10.1111/j.1541-0420.2011.01685.x

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

[10] Einmahl, U. (2009). A new strong invariance principle for sums of independent random vectors. *J. Math. Sci.* **163** 311–327. MR2749123 https://doi.org/10.1007/s10958-009-9676-8

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

[12] Howard, S. R. and Ramdas, A. (2022). Sequential estimation of quantiles with applications to A/B testing and best-arm identification. *Bernoulli* **28** 1704–1728. MR4411508 https://doi.org/10.3150/21-bej1388

[13] Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2020). Time-uniform Chernoff bounds via nonnegative supermartingales. *Probab. Surv.* **17** 257–317. MR4100718 https://doi.org/10.1214/18-PS321

[14] Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *Ann. Statist.* **49** 1055–1080. MR4255119 https://doi.org/10.1214/20-aos1991

[15] Johari, R., Koomen, P., Pekelis, L. and Walsh, D. (2017). Peeking at A/B tests: Why it matters, and what to do about it. In *Proceedings of the* 23rd *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1517–1525.

[16] Johnson, A. and Pollard, T. (2018). sepsis3–mimic. https://doi.org/10.5281/zenodo.1256723

[17] Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A. et al. (2016). MIMIC-III, a freely accessible critical care database. *Sci. Data* **3** 160035.

[18] Kennedy, E. H. (2016). Semiparametric theory and empirical processes in causal inference. In *Statistical Causal Inferences and Their Applications in Public Health Research*. *ICSA Book Ser. Stat.* 141–167. Springer, Berlin. MR3617956

[19] Komlós, J., Major, P. and Tusnády, G. (1975). An approximation of partial sums of independent RV's and the sample DF. I. *Z. Wahrsch. Verw. Gebiete* **32** 111–131. MR0375412 https://doi.org/10.1007/BF00533093

[20] Komlós, J., Major, P. and Tusnády, G. (1976). An approximation of partial sums of independent RV's, and the sample DF. II. *Z. Wahrsch. Verw. Gebiete* **34** 33–58. MR0402883 https://doi.org/10.1007/BF00532688

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

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

[23] LINDEBERG, J. W. (1922). Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Math. Z.* **15** 211–225. MR1544569 https://doi.org/10.1007/BF01494395

[24] MAJOR, P. (1976). The approximation of partial sums of independent RV's. *Z. Wahrsch. Verw. Gebiete* **35** 213–220. MR0415743 https://doi.org/10.1007/BF00532673

[25] MARCINKIEWICZ, J. and ZYGMUND, A. (1937). Sur les fonctions indépendantes. *Fund. Math.* **29** 60–90.

[26] MORROW, G. and PHILIPP, W. (1982). An almost sure invariance principle for Hilbert space valued martingales. *Trans. Amer. Math. Soc.* **273** 231–251. MR0664040 https://doi.org/10.2307/1999203

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

[28] POLLARD, T. J. and JOHNSON, A. E. (2016). The MIMIC-III Clinical Database. https://doi.org/10.13026/C2XW26

[29] POLLEY, E. C. and VAN DER LAAN, M. J. (2010). *Super Learner in Prediction. U.C. Berkeley Division of Biostatistics Working Paper Series* **222**.

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

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

[32] ROBINS, J., LI, L., TCHETGEN, E. and VAN DER VAART, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman. Inst. Math. Stat. (IMS) Collect.* **2** 335–421. IMS, Beachwood, OH. MR2459958 https://doi.org/10.1214/193940307000000527

[33] ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

[34] ROTNITZKY, A., ROBINS, J. M. and SCHARFSTEIN, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J. Amer. Statist. Assoc.* **93** 1321–1339. MR1666631 https://doi.org/10.2307/2670049

[35] SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.* **94** 1096–1120. With comments and a rejoinder by the authors. MR1731478 https://doi.org/10.2307/2669923

[36] SHAHN, Z., SHAPIRO, N. I., TYLER, P. D., TALMOR, D. and LI-WEI, H. L. (2020). Fluid-limiting treatment strategies among sepsis patients in the ICU: A retrospective causal analysis. *J. Crit. Care* **24** 1–9.

[37] SINGER, M., DEUTSCHMAN, C. S., SEYMOUR, C. W., SHANKAR-HARI, M., ANNANE, D., BAUER, M., BELLOMO, R., BERNARD, G. R., CHICHE, J.-D. et al. (2016). The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA* **315** 801–810.

[38] SKOROKHOD, A. (1961). *Research on the Theory of Random Processes*. Kiev Univ., Kiev.

[39] STRASSEN, V. (1964). An invariance principle for the law of the iterated logarithm. *Z. Wahrsch. Verw. Gebiete* **3** 211–226. MR0175194 https://doi.org/10.1007/BF00534910

[40] STRASSEN, V. (1967). Almost sure behavior of sums of independent random variables and martingales. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. II: Contributions to Probability Theory, Part* 1 315–343. Univ. California Press, Berkeley, CA. MR0214118

[41] TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data. Springer Series in Statistics*. Springer, New York. MR2233926

[42] TSYBAKOV, A. B. (2003). Optimal rates of aggregation. In *Learning Theory and Kernel Machines* 303–313. Springer, Berlin.

[43] VAN WALRAVEN, C., AUSTIN, P. C., JENNINGS, A., QUAN, H. and FORSTER, A. J. (2009). A modification of the elixhauser comorbidity measures into a point system for hospital death using administrative data. *Med. Care* **47** 626–633. https://doi.org/10.1097/MLR.0b013e31819432e5

[44] VAN DER LAAN, M. J., POLLEY, E. C. and HUBBARD, A. E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.* **6** Art. 25, 23. MR2349918 https://doi.org/10.2202/1544-6115.1309

[45] VAN DER LAAN, M. J. and ROSE, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data. Springer Series in Statistics*. Springer, New York. MR2867111 https://doi.org/10.1007/978-1-4419-9782-1

[46] VAN DER VAART, A. W., DUDOIT, S. and VAN DER LAAN, M. J. (2006). Oracle inequalities for multi-fold cross validation. *Statist. Decisions* **24** 351–371. MR2305112 https://doi.org/10.1524/stnd.2006.24.3.351

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

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

[49] WAUDBY-SMITH, I., ARBOUR, D., SINHA, R., KENNEDY, E. H. and RAMDAS, A. (2024). Supplement to "Time-uniform central limit theory and asymptotic confidence sequences." https://doi.org/10.1214/24-AOS2408SUPPA, https://doi.org/10.1214/24-AOS2408SUPPB

[50] WAUDBY-SMITH, I. and RAMDAS, A. (2024). Estimating means of bounded random variables by betting. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **86** 1–27. MR4716192 https://doi.org/10.1093/jrsssb/qkad009

[51] YU, M., LU, W. and SONG, R. (2020). A new framework for online testing of heterogeneous treatment effect. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *Vol.* 34 10310–10317.

[52] ZHENG, W. and VAN DER LAAN, M. J. (2010). *Asymptotic Theory for Cross-Validated Targeted Maximum Likelihood Estimation. U.C. Berkeley Division of Biostatistics Working Paper Series* **273**.