Asymptotic Anytime-Valid Inference for U-statistics

Leheng Cai[1*] **Qirui Hu**[2,3,4*†] **Weijia Li**[1*]

Abstract

We study asymptotic anytime-valid confidence sequences for degree-two U-statistics under continuous monitoring. In the nondegenerate case, Hoeffding's projection reduces the problem to a time-uniform central limit theory for the partial sums of the first-order projection, while the canonical remainder is shown to be negligible under mild moment assumptions. A leave-one-out jackknife estimator then yields a fully data-driven procedure, leading to confidence sequences with asymptotic coverage guarantee for the parameter of interest. In the degenerate case, we show that the U-statistic is approximated by a centered quadratic Gaussian-chaos rather than by a simple Gaussian, which poses significant challenges for sequential inference. To address this issue, we novelly develop the Spectrally Allocated Gaussian-chaos Excursion (SAGE) boundary, and then provide plug-in implementations based on truncated spectrum estimation with consistency guarantees. The resulting widths can attain the expected time-uniform optimal rates: $\sqrt{\log\log n/n}$ in the nondegenerate regime and $\log\log n/n$ in the degenerate regime. Several widely used U-statistics are discussed within the proposed framework, and numerical experiments further support the validity of the derived theory.

1 Introduction

U-statistics are a typical device for estimating distributional functionals that can be written as expectations of symmetric kernels. Since Hoeffding's decomposition, they have served as a unifying language for rank statistics, variance and covariance estimation, measures of dispersion and association, and two-sample or independence testing [22, 39, 28, 27, 30, 24, 29]. Their appeal is not only unbiasedness: the same projection calculus gives classical limit theory and separates the regular first-order component from genuinely second-order behavior. In the nondegenerate regime, the first Hoeffding projection reduces the leading term to an average of i.i.d. influence functions, while in the canonical degenerate regime, the projection vanishes and the statistic is instead a quadratic form or Gaussian chaos, with limit laws depending on the spectrum of the associated integral operator [20, 12, 10, 1, 11, 17, 16, 5].

This dichotomy is especially relevant in modern machine learning and nonparametric distributional inference. Kernel maximum mean discrepancy (MMD), Hilbert–Schmidt independence criteria, energy distances, and distance covariance are all naturally expressed through U- or V-statistics; under the null hypothesis of equality or independence, their first-order projections typically vanish, producing the weighted chi-square or Gaussian-chaos null limits [19, 2, 42, 41, 38]. Thus, the degenerate case is not a pathological corner case. It is the operating regime of many distributional tests. At the same time, it is the regime in which calibration is most delicate, because the boundary

*All authors contributed equally and are listed in alphabetical order by surname.
†Corresponding author: huqirui@mail.shufe.edu.cn.
[1]Department of Statistics and Data Science, Tsinghua University.
[2]School of Statistics and Data Science, Shanghai University of Finance and Economics.
[3]Institute of Big Data Research, Shanghai University of Finance and Economics.
[4]Department of Mathematics, Ruhr University Bochum.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

depends on spectral quantities that are usually unknown and must be estimated from the same data stream [26, 37].

For a concrete continuous-monitoring example, suppose factor scores F_i and subsequent returns R_i arrive sequentially in financial quantitive research. Kendall's kernel $h(\boldsymbol{Z}_i, \boldsymbol{Z}_j) = \operatorname{sign}(F_i - F_j)\operatorname{sign}(R_i - R_j)$ with $\boldsymbol{Z}_i = (F_i, R_i)^\top$, gives a nondegenerate U-functional whose confidence sequence can be inspected repeatedly before a data-dependent deployment decision. In the canonical regime, paired returns $\boldsymbol{Z}_i = (X_i, Y_i)^\top$ from two quantitive strategies can be monitored through the MMD kernel $h(\boldsymbol{Z}_i, \boldsymbol{Z}_j) = k(X_i, X_j) + k(Y_i, Y_j) - k(X_i, Y_j) - k(X_j, Y_i)$; under equality of the return distributions its first projection vanishes. However, classical pointwise intervals do not control the probability of a false conclusion over the repeated inspections.

Anytime-valid inference asks for uncertainty statements that remain valid under continuous monitoring and optional stopping. For sample means and related estimating equations, this goal is now well understood through nonnegative martingales, test supermartingales, mixture boundaries, and betting constructions [43, 9, 35, 36, 23, 45, 34]. These methods give finite-sample confidence sequences (CSs) when enough structure is available, for example boundedness, sub-exponential tails, or an exact martingale estimating equation. A complementary line of work develops asymptotic confidence sequences (AsympCSs) by replacing fixed-time central limit approximations with time-uniform strong approximations, thereby recovering CLT-like versatility in sequential settings [44, 18].

For U-statistics, however, the usual mean-based sequential tools do not transfer directly. Adding a new observation changes all pairs involving that observation, so the statistic is not a sum of independent increments. Even in the nondegenerate case, the reduction to a sample mean is only asymptotic and must be uniform over all monitoring times; a fixed-time CLT does not control the event of ever crossing a boundary. One must prove that the canonical Hoeffding remainder is almost surely negligible relative to the iterated-logarithm order, and any plug-in estimation does not disturb time-uniform coverage. Notably, the degenerate case is much harder: there is no Brownian first-order limit, the correct width is of order $\log\log n/n$ rather than $\sqrt{\log\log n/n}$, and a boundary must control a countable quadratic Gaussian chaos with unknown signed eigenvalues.

There is a substantial sequential literature around U-statistics, but it addresses different inferential targets. Classical work studied fixed-width or two-stage sequential confidence procedures and stopping times based on U-statistics [40, 31, 32, 33]. More recent work has used U-statistic processes for retrospective or online change-point detection, including robust Wilcoxon-type monitoring and distributional monitoring based on degenerate U-statistics [13, 14, 25, 3]. These contributions are closely related in motivation, but their calibration is designed for testing change-points or for prescribed stopping rules. They do not provide a general AsympCS for a fixed degree-two U-functional, and they do not give a unified treatment of the nondegenerate and canonical degenerate regimes with data-driven spectral calibration.

This paper fills that gap by developing asymptotic anytime-valid inference for degree-two U-statistics under continuous monitoring. Our central idea is to separate the construction into two components: a probabilistic approximation that holds uniformly over time, and a statistical calibration step that estimates the nuisance quantities appearing in the limiting boundary. This separation yields a modular framework and clarifies precisely where moment conditions, degeneracy, and spectral estimation enter. In the nondegenerate regime, we show that the relevant limiting object is a Brownian partial-sum process associated with the first Hoeffding projection. Based on an almost sure Gaussian approximation and a strongly consistent jackknife variance estimator (Theorems 1 and 2), we construct AsympCSs; see Theorem 3. These sequences connect naturally to classical time-uniform Gaussian boundaries; in particular, the stitched law-of-the-iterated-logarithm boundary achieves the optimal width $\sqrt{\log\log n/n}$.

In the degenerate regime, the limiting object is instead a centered quadratic Gaussian-chaos process. For this case, we derive a Gaussian-chaos approximation and novelly develop the Spectrally Allocated Gaussian-chaos Excursion (SAGE) boundary; see Theorems 4 and 5, which allocates the significance level budget across spectral coordinates and yields an AsympCS with the canonical $\log\log n/n$ scale, matching the law-of-the-iterated-logarithm behavior of canonical degree-two U-statistics. We then provide data-driven plug-in procedures for the degenerate regime based on truncated empirical spectral estimation. Finally, we illustrate the scope of the framework through common examples, including sample variance, Gini mean difference, spatial Kendall's tau and MMD. To our best knowledge, this paper first addressed the asymptotic anytime-valid inference for U-statistics systematically.

The rest of the paper is organized as follows. Section 2 fixes notations, recalls Hoeffding's decomposition, defines AsympCSs, and states the Gaussian boundaries used in the constructions. Section 3 gives the nondegenerate and degenerate theoretical guarantees. Section 5 contains simulation designs and numerical results. Section 6 concludes. Illustrative examples, additional numerical results, and detailed proofs are deferred to the appendix.

2 Preliminaries

Hoeffding decomposition. Let $X_1, \dots, X_n$ be i.i.d. random elements taking values in $\mathcal{X}$. Consider the second-order U-statistic

$$U_n = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} h(X_i, X_j),$$

where $h : \mathcal{X}^2 \to \mathbb{R}$ is a symmetric kernel satisfying $\mathbb{E}h^2(X_1, X_2) < \infty$. Let $\theta = \mathbb{E}\, h(X_1, X_2)$ be the parameter of interest. Define the first-order projection $h_1(x) = \mathbb{E}\{h(x, X_2)\} - \theta$, and the canonical kernel $h_2(x, y) := h(x, y) - \theta - h_1(x) - h_1(y)$. Hoeffding's decomposition gives

$$U_n - \theta = \frac{2}{n} \sum_{i=1}^{n} h_1(X_i) + R_n, \qquad R_n := \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} h_2(X_i, X_j). \tag{1}$$

We call the U-statistic nondegenerate if $\mathrm{Var}(h_1(X_1)) > 0$, and degenerate (canonical) if $h_1(X_1) = 0$ almost surely. For degenerate U-statistics, let $\mathcal{K} : L^2(\mathbb{P}) \to L^2(\mathbb{P})$ be the Hilbert-Schmidt operator $(\mathcal{K}f)(x) = \int \{h(x, y) - \theta\} f(y) d\mathbb{P}(y)$. Then $\mathcal{K}$ is compact and self-adjoint, and admits the spectral expansion in $L^2(\mathbb{P} \times \mathbb{P})$, i.e. for the eigensystems $(\lambda_\ell, \psi_\ell)_{\ell \ge 1}$ of $\mathcal{K}$,

$$h(x, y) - \theta = \sum_{\ell \ge 1} \lambda_\ell \psi_\ell(x) \psi_\ell(y). \tag{2}$$

Asymptotic anytime-valid sequential inference. In this paper, we adopt the following framework of asymptotic anytime-valid inference [44].

Definition 1 (AsympCSs)**.** *Let $\mathcal{T}$ be a totally ordered infinite set. The intervals $\left[\widehat{\theta}_t - \ell_t, \widehat{\theta}_t + u_t\right]_{t \in \mathcal{T}}$ with $\ell_t, u_t > 0$ form an $(1-\alpha)$-AsympCS for a sequence of real parameters $\{\theta_t\}_{t \in \mathcal{T}}$ if there exists a (typically unknown) non-asymptotic $(1-\alpha)$-CS $\left[\widehat{\theta}_t - \ell_t^\star, \widehat{\theta}_t + u_t^\star\right]_{t \in \mathcal{T}}$ such that $\mathbb{P}\left(\forall t \in \mathcal{T}, \theta_t \in \left[\widehat{\theta}_t - \ell_t^\star, \widehat{\theta}_t + u_t^\star\right]\right) \geqslant 1 - \alpha$, and $\ell_t^\star / \ell_t \xrightarrow{a.s.} 1$, $u_t^\star / u_t \xrightarrow{a.s.} 1$.*

Definition 2 (Asymptotic time-uniform coverage)**.** *For each $m \in \mathbb{N}$, let $\{C_n(m)\}_{n \ge m}$ be a sequence of random sets. We say that this delayed-start family has asymptotic time-uniform $(1-\alpha)$-coverage for θ if*

$$\liminf_{m \to \infty} \mathbb{P}(\theta \in C_n(m) \text{ for every } n \ge m) \ge 1 - \alpha.$$

As discussed in Section 2.1 of [44], however, a probability guarantee of the form given in Definition 2 cannot be adopted directly as the definition of an AsympCS. Indeed, for a nonasymptotic confidence sequence constructed at an arbitrary level α' (even if $\alpha' \gg \alpha$), the eventual coverage probability converges to one as $m \to \infty$. Consequently, the corresponding inequality would hold vacuously and would fail to distinguish between different nominal confidence levels.

Unlike existing nonasymptotic inference results for means of random variables, which often rely on strong boundedness or moment-generating-function assumptions [45, 23], this asymptotic framework allows us to establish universal closed-form boundaries for both nondegenerate and degenerate U-statistics under mild moment conditions.

On the other hand, a central tool for constructing AsympCSs for the mean of i.i.d. random variables is a time-uniform boundary for Gaussian partial sums [44], of which the formal definition is given below.

Definition 3 (Gaussian boundary)**.** *Fix $\alpha \in (0,1)$. A map $\gamma_{\alpha,m} : \mathbb{N} \to (0,\infty)$ is called a $(1-\alpha)$ Gaussian boundary starting from m if, for i.i.d. standard Gaussian random variables $\{Z_i\}_{i\in\mathbb{N}}$,*

$$\mathbb{P}\left(\forall n \geq m : \left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right| \leq \gamma_{\alpha,m}(n)\right) \geq 1-\alpha. \tag{3}$$

The starting point m marks the end of the cold-start phase and should be specified before monitoring. Its role is analogous to the sample size in the classical asymptotic approximation: increasing m gives the Gaussian and plug-in errors more time to become negligible, while the formal delayed-start guarantees are obtained as $m \to \infty$. Several standard choices of the boundary $\gamma_{\alpha,m}(\cdot)$ are available for non-asymptotic CSs of Gaussian means in the literature. A representative example is the stitched boundary of [23], which achieves the concentration rate $\mathcal{O}\left(\sqrt{\log\log n/n}\right)$:

$$\gamma^{\mathrm{LIL}}_{\alpha,m}(n) = \frac{\eta^{1/4}+\eta^{-1/4}}{\sqrt{2n}}\sqrt{s\log\log\left(\max\left\{\frac{\eta n}{m}, e\right\}\right) + \log\frac{\zeta(s)}{\alpha(\log\eta)^s}}, \tag{4}$$

where $\eta, s > 1$, and $\zeta(\cdot)$ is the Riemann zeta function.

Another classical construction is Robbins' Gaussian mixture boundary [35, 36], which attains the rate $\mathcal{O}(\sqrt{\log n/n})$:

$$\gamma^{\mathrm{GM}}_{\alpha,m}(n) = \sqrt{\left[\{g^{-1}(\alpha)\}^2 + \log(n/m)\right]/n}, \quad \text{where } g(a) = 2\{1-\Phi(a)+a\phi(a)\}. \tag{5}$$

Here, $\Phi(\cdot)$ and $\phi(\cdot)$ denote the CDF and PDF of a standard Gaussian distribution, respectively. These bounds will be used as building blocks for our proposed anytime-valid inference procedure for U-statistics.

3 Main results

Non-degenerate case. We first consider that the kernel is non-degenerate at the first order, i.e., $\sigma^2 := \mathrm{Var}\{h_1(X_1)\} > 0$.

The Hoeffding decomposition in (1) implies that the leading behavior of a nondegenerate second-order U-statistic is governed by an average of i.i.d. first-order projections. By establishing an almost sure Gaussian approximation for the first Hoeffding projection with positive variance and showing that the canonical remainder is asymptotically negligible almost surely, Theorem 1 rigorously shows that the nondegenerate U-statistic can be approximated almost surely by the mean of a sequence of i.i.d. Gaussian random variables.

Theorem 1. *Suppose that $\mathbb{E}|h(X_1,X_2)|^{2+\delta} < \infty$ for some $\delta > 0$, and that the kernel is non-degenerate at the first order, i.e., $\sigma^2 > 0$. On an enriched probability space there exists i.i.d. standard Gaussian random variables $Z_1,\ldots,Z_n$ such that*

$$\left|U_n - \theta - \frac{2\sigma}{n}\sum_{i=1}^{n} Z_i\right| = o_{a.s.}\left(n^{-1+1/(2+\delta)}\right).$$

We consider the jackknife-type variance estimator for the unknown variance σ^2 in Theorem 1,

$$\widehat{\sigma}^2_n = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{n-1}\sum_{j\neq i} h(X_i,X_j)\right\}^2 - U_n^2. \tag{6}$$

The following Theorem 2 establishes the strong consistency of $\widehat{\sigma}^2_n$.

Theorem 2. *Suppose that $\mathbb{E}|h(X_1,X_2)|^4 < \infty$. The variance estimator $\widehat{\sigma}^2_n$ defined in (6) is strongly consistent: $\left|\widehat{\sigma}^2_n - \sigma^2\right| = o_{a.s.}\left(\sqrt{\log\log n/n}\right)$.*

Strong consistency of related jackknife variance estimators is classical; see, for example, [40]. The contribution of Theorem 2 used here is the explicit almost-sure rate, vital for the establishing the asymptotic time-uniform coverage of AsympCSs.

For classical asymptotic inference for non-degenerate U-statistics, one typically applies the central limit theorem and uses non-asymptotic standard normal quantiles as critical values to construct asymptotic confidence intervals (CIs). Motivated by Theorem 1 and the analogy with classical asymptotic inference, we use Gaussian boundaries to construct AsympCSs. The following Theorem 3 provides the corresponding theoretical justification.

Theorem 3. *Suppose that* $\mathbb{E}|h(X_1, X_2)|^4 < \infty$. *Then,* $[U_n \pm 2\widehat{\sigma}_n \gamma_{\alpha,m}(n)]_{n\geq m}$ *forms an* $(1-\alpha)$*-AsympCS for* θ *starting from* m. *Furthermore,*

$$\liminf_{m\to\infty} \mathbb{P}\left\{\forall n \geq m : \theta \in [U_n \pm 2\widehat{\sigma}_n \gamma_{\alpha,m}(n)]\right\} \geq 1-\alpha. \tag{7}$$

By plugging in the Gaussian boundaries in (4) or (5), we obtain universal AsympCSs for nondegenerate U-statistics. In particular, the plug-in boundary based on (4) achieves the optimal rate, since the law of the iterated logarithm for nondegenerate U-statistics yields

$$\limsup_{n\to\infty} \sqrt{n/(2\log\log n)}|U_n - \theta| = 2\sigma, \quad \text{a.s.}$$

In addition, we note that as long as σ^2 admits a polynomial-rate estimator rather than merely an arbitrary $\mathcal{O}_{a.s.}(1)$-consistent estimator, the resulting AsympCS enjoys the asymptotic time-uniform coverage guarantee in (7).

Degenerate case. We now consider the canonical case, where $h_1(X_1) = 0$ almost surely. In this regime, the U-statistic is no longer dominated by the first-order projection in the Hoeffding decomposition; instead, it is entirely the second-order degenerate component. Consequently, its asymptotic behavior changes significantly.

The following Theorem 4 shows that the degenerate U-statistic admits an almost sure approximation by a centered quadratic Gaussian chaos process.

Theorem 4. *Assume that* $\mathbb{E}\left\{h(X,X')\log|h(X,X')|\right\}^2 + \mathbb{E}\left\{h(X,X)\log|h(X,X)|\right\}^2 < \infty$ *and the kernel is degenerate at the first order, i.e.,* $\sigma^2 = 0$. *On an enriched probability space there exist independent standard Brownian motions* $\{W_\ell(\cdot)\}_{\ell\geq 1}$ *such that*

$$\left| U_n - \theta - n^{-2}\sum_{\ell\geq 1} \lambda_\ell \left\{W_\ell^2(n) - n\right\} \right| = \mathcal{O}_{a.s.}\left(\log\log n/n\right), \tag{8}$$

where $\{\lambda_\ell\}_{\ell\geq 1}$ *is the eigenvalues defined in* (2).

Let $(\beta_{\ell,n})_{\ell\geq 1}$ be a nonnegative weight array with $\sum_{\ell\geq 1}\beta_{\ell,n} = 1$, and we write β_ℓ for $\beta_{\ell,n}$ for simplicity. Let $\Lambda = \sum_{\ell\geq 1}\lambda_\ell$, $\Lambda^+ = \sum_{\ell:\lambda_\ell>0}\lambda_\ell$, $\Lambda^+_\beta = \sum_{\ell:\lambda_\ell>0}\lambda_\ell\log(1/\beta_\ell)$ and $\Lambda^+_{\beta,g} = \sum_{\ell:\lambda_\ell>0}\lambda_\ell\left\{g^{-1}(\alpha\beta_\ell)\right\}^2$, where $g(a) = 2\{1-\Phi(a)+a\phi(a)\}$. We now introduce the Spectrally Allocated Gaussian-chaos Excursion (SAGE) boundary $\Upsilon_{\alpha,m}(\cdot)$, for controlling quadratic Gaussian-chaos processes. For a fixed significance level $\alpha \in (0,1)$, we distribute the budget α across the Brownian coordinates by setting $\alpha_\ell = \alpha\beta_\ell$, apply a Gaussian mean boundary to each $W_\ell(\cdot)$, combine the coordinatewise guarantees via a union bound, and aggregate the resulting bounds with the spectral weights λ_ℓ. This yields the following time-uniform upper boundary for the centered quadratic Gaussian-chaos process.

Theorem 5. *Suppose that* $\sum_{\ell\geq 1}|\lambda_\ell| < \infty$, $\Lambda^+ > 0$, $\Lambda^+_\beta < \infty$ *and* $\{W_\ell(\cdot)\}_{\ell\geq 1}$ *are independent standard Brownian motions. For any* $\alpha \in (0,1)$, $m \geq 1$,

$$\mathbb{P}\left(\forall n \geq m :\ n^{-2}\sum_{\ell\geq 1}\lambda_\ell\left\{W_\ell^2(n) - n\right\} \leq \Upsilon_{\alpha,m}(n)\right) \geq 1-\alpha,$$

where $\Upsilon_{\alpha,m}(n)$ *can be chosen in either of the following two ways:*

$$\Upsilon^{LIL}_{\alpha,m}(n) = \frac{\left(\eta^{1/4}+\eta^{-1/4}\right)^2}{2n}\left[\left\{s\log\log\left(\max\left\{\frac{\eta n}{m}, e\right\}\right) + \log\frac{\zeta(s)}{\alpha(\log\eta)^s}\right\}\Lambda^+ + \Lambda^+_\beta\right] - \frac{\Lambda}{n}, \tag{9}$$

$$\Upsilon^{GM}_{\alpha,m}(n) = n^{-1}\left[\Lambda^+\log\left(n/m\right) + \Lambda^+_{\beta,g} - \Lambda\right]. \tag{10}$$

When applying the Gaussian boundaries in (4) and (5) coordinatewise to each $W_\ell(\cdot)$, we obtain two variants of the SAGE boundary, (9) and (10), with widths of order $\mathcal{O}(\log\log n/n)$ and $\mathcal{O}(\log n/n)$, respectively. The rate of $\Upsilon_{\alpha,m}(\cdot)$ in (9) matches the law of the iterated logarithm rate [17], since $\limsup_{n\to\infty} n|U_n-\theta|/\log\log n<\infty$ almost surely for canonical U-statistics.

Moreover, we note that the SAGE boundary proposed in Theorem 5 is a one-sided upper boundary. This is sufficient for many classical statistical inference based on degenerate U-statistics, including distributional tests based on MMD or energy distance, where evidence against the null is typically accumulated in the upper tail. We defer the lower boundary results to Section C.1 and focus on the upper boundary throughout the paper and the subsequent example.

In practice, implementation of the proposed SAGE boundaries requires consistent eigenvalue estimation, which we formalize in Assumption 1.

Assumption 1. *For a small $\iota>0$, there exist consistent estimators $\widehat{\Lambda}$, $\widehat{\Lambda}^+$, $\widehat{\Lambda}^+_\beta$ and $\widehat{\Lambda}^+_{\beta,g}$ such that $|\widehat{\Lambda}-\Lambda|+|\widehat{\Lambda}^+-\Lambda^+|+|\widehat{\Lambda}^+_\beta-\Lambda^+_\beta|+|\widehat{\Lambda}^+_{\beta,g}-\Lambda^+_{\beta,g}|=\mathcal{O}_{a.s.}(n^{-\iota})$.*

Assumption 1 is a high level condition on the estimated spectral functionals. Proposition C.2 in the appendix provides empirical estimators and sufficient conditions that guarantee Assumption 1.

Theorem 6. *Under Assumption 1 and the conditions in Theorems 4–5, for any $\alpha\in(0,1)$ and $m\geq 1$, the sequence $\left[U_n-\widehat{\Upsilon}_{\alpha,m}(n),\infty\right]_{n\geq m}$ is an $(1-\alpha)$-AsympCS for θ in the sense of Definition 1. In addition, it satisfies the asymptotic time-uniform coverage in Definition 2:*

$$\liminf_{m\to\infty}\mathbb{P}\left(\forall n\geq m:\ U_n\leq\theta+\widehat{\Upsilon}_{\alpha,m}(n)\right)\geq 1-\alpha,$$

where $\widehat{\Upsilon}_{\alpha,m}(n)$ can be chosen in either of the following two ways:

$$\widehat{\Upsilon}^{LIL}_{\alpha,m}(n)=\frac{\left(\eta^{1/4}+\eta^{-1/4}\right)^2}{2n}\left[\left\{s\log\log\left(\max\left\{\frac{\eta n}{m},e\right\}\right)+\log\frac{\zeta(s)}{\alpha(\log\eta)^s}\right\}\widehat{\Lambda}^++\widehat{\Lambda}^+_\beta\right]-\frac{\widehat{\Lambda}}{n},\tag{11}$$

$$\widehat{\Upsilon}^{GM}_{\alpha,m}(n)=\frac{1}{n}\left[\widehat{\Lambda}^+\log\left(\frac{n}{m}\right)+\widehat{\Lambda}^+_{\beta,g}-\widehat{\Lambda}\right].\tag{12}$$

Theorem 6 provides the asymptotic coverage guarantee for the plug-in SAGE boundaries. The duality between CSs and sequential tests makes our framework well adapted for modern sequential testing and continuous monitoring. Formally, an asymptotic $(1-\alpha)$-CS $\{\mathcal{C}_n\}_{n\geq m}$ is equivalent to a family of asymptotic level-α sequential tests for $H_0:\theta=\theta_0$. By defining the rejection rule as $\phi_n=\mathbf{1}\{\theta_0\notin\mathcal{C}_n\}$, the Type I error is asymptotically controlled at level α for any stopping time.

Implementation. Building upon theoretical guarantees established for both regimes, we provide two algorithmic templates for constructing these anytime-valid AsympCSs below.

Non-degenerate case. For $r_{i,n}:=\sum_{j\leq n,j\neq i}h(X_i,X_j)$,

$$U_n=\frac{1}{n(n-1)}\sum_{i=1}^n r_{i,n},\qquad \widehat{\sigma}_n^2=\frac{1}{n(n-1)^2}\sum_{i=1}^n r_{i,n}^2-U_n^2.$$

When X_{n+1} arrives, let $D_{n+1}=\sum_{i=1}^n h(X_i,X_{n+1})$. Then

$$\begin{aligned}U_{n+1}&=\frac{n-1}{n+1}U_n+\frac{2D_{n+1}}{n(n+1)},\quad r_{i,n+1}=r_{i,n}+h(X_i,X_{n+1}),\ i\leq n,\\ \widehat{\sigma}^2_{n+1}&=\frac{\sum_{i=1}^n\{r_{i,n}+h(X_i,X_{n+1})\}^2+D_{n+1}^2}{(n+1)n^2}-U_{n+1}^2,\quad r_{n+1,n+1}=D_{n+1}.\end{aligned}\tag{13}$$

Thus, a general kernel requires n new kernel evaluations, $\mathcal{O}(n)$ arithmetic, and $\mathcal{O}(n)$ row-sum storage per observation, rather than recomputation from scratch; see Algorithm 1.

Algorithm 1 Nondegenerate AsympCSs

Require: Data stream $X_1, X_2, \ldots$; symmetric kernel h; level $\alpha \in (0,1)$; delayed start $m \geq 2$.
Ensure: Two-sided $(1-\alpha)$-AsympCSs $(\mathcal{C}_n^{\mathrm{ND}})_{n \geq m}$ for $\theta = \mathbb{E}h(X_1, X_2)$.
1: **for** $n = m, m+1, \ldots$ **do**
2: Compute $U_n, \widehat{\sigma}_n^2$ using (13).
3: Compute boundary $\gamma_{\alpha,m}(\cdot)$ using (4) or (5).
4: Output $\mathcal{C}_n^{\mathrm{ND}} \leftarrow [U_n - 2\widehat{\sigma}_n \gamma_{\alpha,m}(n),\ U_n + 2\widehat{\sigma}_n \gamma_{\alpha,m}(n)]$.
5: **end for**

Algorithm 2 Degenerate AsympCSs

Require: Data stream $X_1, X_2, \ldots$; canonical symmetric kernel h; level $\alpha \in (0,1)$; delayed start $m \geq 2$; truncation number L_n; weights $(\beta_\ell)_{\ell \geq 1}$ with $\sum_{\ell \geq 1} \beta_\ell = 1$
Ensure: One-sided $(1-\alpha)$-AsympCSs $(\mathcal{C}_n^{\mathrm{D}})_{n \geq m}$ for $\theta = \mathbb{E}h(X_1, X_2)$.
1: **for** $n = m, m+1, \ldots$ **do**
2: Compute U_n using (13).
3: Form the centered Gram matrix $\widehat{\mathbf{K}}_n \leftarrow h(X_i, X_j) - U_n, 1 \leq i, j \leq n$.
4: Retain the top L_n eigenvalues of the scaled matrix $n^{-1}\widehat{\mathbf{K}}_n$ in absolute value: $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{L_n}$.
5: Compute $\widehat{\Lambda}$, $\widehat{\Lambda}^+$, $\widehat{\Lambda}^+_\beta$, $\widehat{\Lambda}^+_{\beta,g}$ provided in Proposition C.2.
6: Compute the SAGE boundary $\widehat{\Upsilon}_{\alpha,m}(n)$ using (11) or (12).
7: Output $\mathcal{C}_n^{\mathrm{D}} \leftarrow [U_n - \widehat{\Upsilon}_{\alpha,m}(n), \infty)$.
8: **end for**

Degenerate case. A key practical consideration in implementing Algorithm 2 is the choice of the significance-level allocation weights $(\beta_\ell)_{\ell \geq 1}$. To tighten the entropy term in the SAGE boundary, a natural choice is to minimize Λ^+_β subject to $\beta_\ell \geq 0$ and $\sum_{\ell \geq 1} \beta_\ell = 1$. Suppose that $\Lambda^+ > 0$ and let $p_\ell = (\lambda_\ell)_+ / \Lambda^+$ where $(\lambda_\ell)_+ := \max\{\lambda_\ell, 0\}$. Then

$$\Lambda^+_\beta = \sum_{\ell \geq 1} (\lambda_\ell)_+ \log(1/\beta_\ell) = \Lambda^+ \left\{ \sum_{\ell \geq 1} p_\ell \log(1/p_\ell) + \mathrm{KL}(\boldsymbol{p} \| \boldsymbol{\beta}) \right\},$$

where the KL divergence $\mathrm{KL}(\boldsymbol{p}\|\boldsymbol{\beta}) := \sum_{\ell \geq 1} p_\ell \log(p_\ell/\beta_\ell)$ is nonnegative and vanishes if and only if $\beta_\ell = p_\ell$ for all $\ell \geq 1$. Hence, the corresponding oracle allocation is $\beta^\star_\ell = (\lambda_\ell)_+ / \Lambda^+$.

Since the population spectrum is unknown, we use its empirical analogue $\widehat{\beta}_{\ell,n} = (\widehat{\lambda}_{\ell,n})_+ / \widehat{\Lambda}^+$ for $\ell \leq L_n$, where $\widehat{\Lambda}^+ = \sum_{\ell \leq L_n} (\widehat{\lambda}_{\ell,n})_+$. at each $n = m, m+1, \ldots$. Unlike the fixed allocation used in the preceding theoretical development, we note that $\widehat{\beta}_{\ell,n}$ is data dependent. Proposition C.3 in the appendix further shows that the data-driven spectral quantities satisfy Assumption 1, thus it provides the theoretical justification for the data-driven allocation used in Algorithm 2: although the weights $\widehat{\beta}_{\ell,n}$ are updated from the observed spectrum, the asymptotic validity result of Theorem 6 still applies.

4 Examples

Here, we introduce four concrete examples, which are adopted in numerical experiments.

Sample variance. The target parameter $\theta = \mathrm{Var}(X_1)$ is estimated by the kernel: $h(x,y) = (x-y)^2/2$. Let $\mu = \mathbb{E}X_1$. The first-order projection and its variance are given by:

$$h_1(x) = \left\{(x-\mu)^2 - \theta\right\}/2, \quad \sigma^2 = \mathrm{Var}\left\{(X_1 - \mu)^2\right\}/4.$$

If $X_1 \sim \mathcal{N}(\mu, \theta)$, the fourth central moment is $3\theta^2$, which yields a closed-form variance $\sigma^2 = \theta^2/2$. Since Theorem 2 requires $\mathbb{E}h^4 < \infty$, the sample variance mandates a eighth-moment condition on X_1, which explains its mild sensitivity to heavier-tailed data observed in our simulations.

Gini mean difference (GMD). The target $\theta = \mathbb{E}|X_1 - X_2|$ is associated with the kernel: $h(x, y) = |x - y|$. The first-order projection and its corresponding variance are defined as:

$$h_1(x) = \mathbb{E}|x - X_2| - \theta, \quad \sigma^2 = \mathrm{Var}\{\mathbb{E}(|X_1 - X_2| \mid X_1)\}.$$

If $X_1 \sim \mathcal{N}(\mu, v)$, the target evaluates to $\theta = 2\sqrt{v/\pi}$, and the projection variance admits the exact form $\sigma^2 = v\left\{1/3 + (2\sqrt{3} - 4)/\pi\right\}$. A finite fourth moment on X_1 guarantees the asymptotic validity, making GMD more robust to heavy tails.

Spatial Kendall's tau. Consider bivariate data $\boldsymbol{X}_i = (X_{i1}, X_{i2})^\top \in \mathbb{R}^2$. The kernel for the off-diagonal entry of the two-dimensional spatial Kendall's tau matrix [6] is:

$$h(\boldsymbol{x}, \boldsymbol{y}) = \frac{(x_1 - y_1)(x_2 - y_2)}{\|\boldsymbol{x} - \boldsymbol{y}\|_2^2} \mathbf{1}\{\boldsymbol{x} \neq \boldsymbol{y}\}.$$

The target $\theta = \mathbb{E}\{h(\boldsymbol{X}_1, \boldsymbol{X}_2)\} \in [-1/2, 1/2]$ provides a robust measure of directional association between two coordinates. Under a bivariate elliptical distribution with shape correlation $\rho \in [-1, 1]$, the target has a closed form $\theta = \left(1 - \sqrt{1 - \rho^2}\right)/(2\rho)$ when $\rho \neq 0$. In our simulation, we set $\rho = 0.6$, which then gives $\theta = 1/6$. Further details are deferred to Appendix A.4. Since the kernel is bounded, all moment conditions required by Theorem 3 hold automatically, which explains the stable coverage observed in the heavy-tailed settings considered here.

MMD with Gaussian kernel. Let $\boldsymbol{Z}_i = (X_i, Y_i)^\top$ denote a paired observation where $X_i \sim P$ and $Y_i \sim Q$. The unbiased squared MMD statistic [19] employs the kernel:

$$h(\boldsymbol{z}, \boldsymbol{z}') = k(x, x') + k(y, y') - k(x, y') - k(x', y),$$

where $k(x, y) = \exp(-|x - y|^2/2)$ is the Gaussian kernel. The target parameter is $\theta = \mathbb{E}h(\boldsymbol{Z}_1, \boldsymbol{Z}_2) =: \mathrm{MMD}^2(P, Q)$. The first-order projection is defined generally as $h_1(\boldsymbol{z}) = \mathbb{E}h(\boldsymbol{z}, \boldsymbol{Z}) - \theta$. Under the null hypothesis that $P = Q$, a direct calculation yields $\theta = 0$ and the projection vanishes: $\sigma^2 = \mathrm{Var}\{h_1(\boldsymbol{Z})\} = 0$, then the U-statistic completely loses its first-order term and falls exactly into the canonical degenerate case.

5 Simulations

Non-degenerate case. For non-degenerate U-statistics, we consider GMD under three data generating distributions: the standard Gaussian, t_{10} and Laplace with density $f(x) = \exp\left(-\sqrt{2}|x|\right)/\sqrt{2}$. To evaluate the empirical coverage and the average width of the proposed AsympCSs, we conduct Monte Carlo simulations with 500 independent replications. We set the nominal error rate to $\alpha = 0.05$ throughout, and monitor the U-statistics continuously from a start $m = 400$ up to $n_{\max} = 10,000$.

We compare against the baseline **classical pointwise CIs** based on the asymptotic normality and evaluate two variants of the AsympCS in (7): **AsympCS-LIL** using the stitched boundary in (4) with $\eta = 2.0$ and $s = 1.4$ suggested by [23], and **AsympCS-GM** using the Gaussian-mixture boundary in (5). The results are illustrated in Figures 1 and A.2-A.3 in Appendix A.2. Additional simulation results for other U-statistics, including sample variance and Spatial Kendall's tau, are deferred to Appendix A.4. As shown in the left panel of Figure 1, the classical CIs suffer from a severe loss of coverage due to repeated sequential evaluations; their cumulative miscoverage rate rapidly climbs above the nominal α level. In contrast, both AsympCS-GM and AsympCS-LIL maintain the miscoverage rate below α uniformly across the entire horizon. The plot of one sample path for U_n in the right panel of Figure 1 further illustrates how the valid AsympCSs safely envelop the true parameter.

The middle panel of Figure 1 traces the average half-widths of the proposed AsympCSs for GMD under the standard Gaussian distribution. As expected from theory, the AsympCSs are wider than the classical pointwise CIs, reflecting the necessary correction for continuous monitoring. Notably, although the AsympCS-GM boundary shrinks at the slower rate $\sqrt{\log n/n}$, it yields sharper intervals in the finite-sample situation. This observation is consistent with [23]: the GM boundary can be tighter at moderate sample sizes, but is eventually outperformed by the AsympCS-LIL boundary, which attains the asymptotically optimal $\sqrt{\log\log n/n}$ rate. A similar phenomenon also emerges

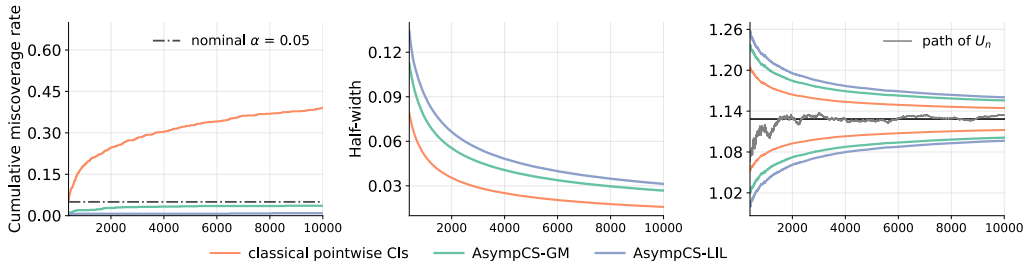


Figure 1: Left: cumulative miscoverage rates of the proposed AsympCSs and the classical pointwise CIs for GMD under the standard Gaussian distribution. Middle: averaged half-widths of the proposed AsympCSs and the classical pointwise CIs. Right: a single sample path of the statistics U_n for GMD alongside the three boundaries, where the black horizontal line indicates the true parameter. The horizontal axis in all three panels represents the sample size.

in our degenerate experiments, translating into higher finite-sample power using the SAGE-GM boundary.

Also we consider the sensitivity to the cold-start parameter m. We examine the reliability of the asymptotic coverage under continuous monitoring by varying the cold-start parameter $m \in \{50, 100, 200\}$. Figure 2 reports the cumulative miscoverage rates of the AsympCS-LIL for GMD under the standard Gaussian and the t_{10} distributions. In both settings, once the start time $m \geq 50$, the cumulative miscoverage rate remains below the level α for up to 10^4 observations. It shows that a moderate cold-start absorbs early finite-sample deviations.

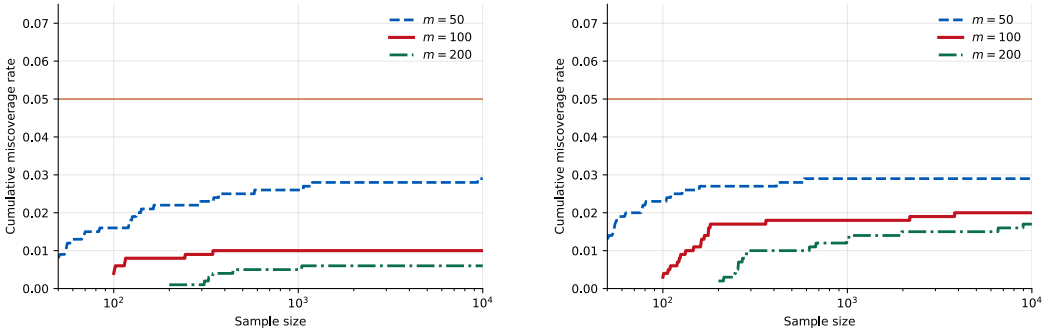


Figure 2: Sensitivity of AsympCS-LIL for GMD to the cold-start parameter $m \in \{50, 100, 200\}$. Cumulative miscoverage rates are reported under the standard Gaussian (left panel) and t_{10} (right panel) distributions.

Degenerate case. For degenerate U-statistics, we study two-sample sequential testing based on MMD with Gaussian kernel over the time horizon from $m = 400$ to $n_{\max} = 2{,}000$. The truncation number L_n is set to be $\lfloor n^{1/4} \rfloor$. We evaluate the sequential MMD test under the standard Gaussian null $H_0 : P = Q$ and under mean-shift alternatives $H_1 : P \neq Q$, parameterized by δ. We compare against the baseline **classical test** using quantiles of centered weighted sum of chi-squared random variables as critical values, and employ the proposed SAGE boundaries in (11) and (12), denoted by **SAGE-LIL** and **SAGE-GM** respectively. All results are based on 500 independent Monte Carlo replications. Additional simulation results under other null distributions are deferred to Appendix A.3.

As shown in the left panel of Figure 3, the classical testing rule completely fails under continuous monitoring, with its empirical rejection rate rapidly accumulating beyond the nominal α level. In contrast, the SAGE boundaries yield sizes close to zero under H_0, while their power approaches one as the sample size increases under H_1 where $\delta = 0.3$ in the simulation. It confirms that our framework achieves rigid Type I error control and have enough power simultaneously. The right panel of Figure 3 also visualizes the underlying mechanism via a single sample path under the null.

To examine the sensitivity to signal strength, we evaluate the power over a grid of mean shifts $\delta \in [0, 0.45]$ at the fixed sample size $n = 2000$. As depicted in the middle panel of Figure 3, the empirical power based on our SAGE boundaries increases from zero to one, demonstrating their effectiveness for sequential testing.

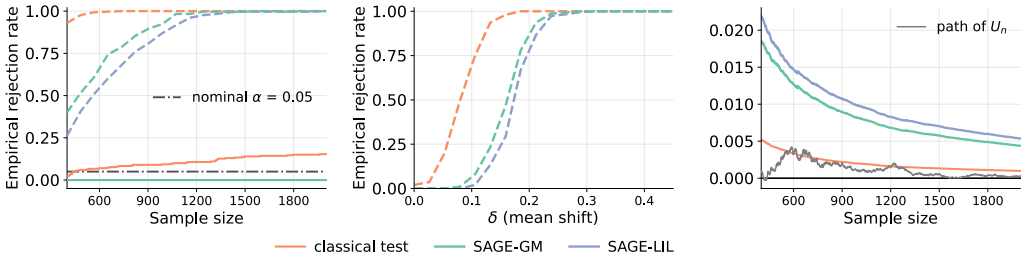


Figure 3: Left: the size and power comparison of the proposed sequential testing procedure using the SAGE boundaries and the classical test procedure for sequential two-sample test with MMD kernel statistics under standard Gaussian distribution. Dashed lines represent the power under H_1 with $\delta = 0.3$, while solid lines represent the size under H_0 with $\delta = 0$. Middle: the empirical power over mean shifts $\delta \in [0, 0.45]$ at sample size $n = 2000$. Right: a single sample path of the MMD test statistic U_n alongside the three boundaries under H_0.

The empirical performance of the sequential test relies on the significance level budget allocation across the infinite eigenvalue spectrum. We compared various allocation strategies, including polynomial decay weights $\beta_\ell = \ell^{-b}/\zeta(b)$, exponential decay weights $\beta_\ell = (1 - e^{-c})e^{-c(\ell-1)}$, and our data-driven weights $\beta_\ell \propto \max\left\{\widehat{\lambda}_\ell, 0\right\}/\widehat{\Lambda}^+$, as shown in Figure A.1 in Appendix A.1. These results highlight theoretical role of $\log(1/\beta_\ell)$: a larger b concentrates the budget on the leading eigenvalues and thus widens the boundary, while a larger c inflates the boundary through the rapid tail growth of $\log(1/\beta_\ell)$. Remarkably, the proposed data-driven procedure optimally distributes the budget, yielding the tightest boundary while remaining below the nominal level. This justifies our adoption of this data-driven approach to compute β_ℓ throughout the degenerate sequential testing experiments.

6 Concluding remarks

This paper develops asymptotic anytime-valid inference for degree-two U-statistics under continuous monitoring. In the nondegenerate regime, Hoeffding's decomposition reduces the problem to the first-order projection, leading to AsympCS with LIL scale $\sqrt{\log\log n/n}$. In the degenerate regime, the limit is instead a spectral quadratic Gaussian chaos with unknown signed eigenvalues. The proposed SAGE boundary separates probability calibration from spectral estimation and yields a one-sided plug-in AsympCSs with the sharper $\log\log n/n$ scale under a weighted spectral approximation condition, covering degenerate kernel and distance-based tests.

Several limitations remain. The theory is restricted to scalar degree-two U-statistics; vector- or matrix-valued extensions require simultaneous time-uniform approximations and control of dimension-dependent errors, especially for spatial Kendall matrices and high-dimensional elliptical models [6, 15, 21]. The algorithms are not yet streaming-optimized, since exact updates, pairwise storage, and repeated eigendecompositions are costly; scalable versions may need incomplete U-statistics, incremental summaries, randomized low-rank updates, or linear-time kernel approximations [19, 46, 7]. Finally, the degenerate theory assumes a verified spectral approximation rate, and finite-sample cold starts, local degeneracy, and automatic switching between regimes remain open.

Acknowledgments and Disclosure of Funding

Qirui Hu's research was supported by National Natural Science Foundation of China (NSFC) (Grant Nos.12601520), the Shanghai Engineering Research Center of Finance Intelligence (Grant No. 19DZ2254600) and by TRR 391 *Spatio-temporal Statistics for the Transition of Energy and Transport* (Project number 520388526) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation).

The authors declare no competing interests.

References

[1] Miguel A. Arcones and Evarist Giné. Limit theorems for U-processes. *The Annals of Probability*, 21(3):1494–1542, 1993.

[2] Ludwig Baringhaus and Carsten Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206, 2004.

[3] Cooper Boniece, Lajos Horvath, and Lorenzo Trapani. Sequential monitoring for distributional changepoint using degenerate U-statistics, 2025.

[4] Chris Brislawn. Traceable integral kernels on countably generated measure spaces. *Pacific Journal of Mathematics*, 150(2):229–240, 1991.

[5] Leheng Cai and Qirui Hu. Approximation theorems for high-dimensional canonical u-statistics: Gaussian chaos and phase transition. *arXiv preprint arXiv:2609.20529*, 2026.

[6] Kyungmee Choi and John I. Marden. A multivariate version of kendall's tau. *Journal of Nonparametric Statistics*, 9(3):261–293, 1998.

[7] Kacper Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems 28*, 2015.

[8] Miklós Csörgő and Pál Révész. *Strong Approximations in Probability and Statistics*. Academic Press, New York, 1981.

[9] D. A. Darling and Herbert Robbins. Iterated logarithm inequalities. *Proceedings of the National Academy of Sciences of the United States of America*, 57(5):1188–1192, 1967.

[10] Piet de Jong. A central limit theorem for generalized quadratic forms. *Probability Theory and Related Fields*, 75(2):261–277, 1987.

[11] Victor H. de la Peña and Evarist Giné. *Decoupling: From Dependence to Independence*. Springer, New York, 1999.

[12] Herold Dehling, Manfred Denker, and Walter Philipp. Invariance principles for von Mises and U-statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 67(2):139–167, 1984.

[13] Herold Dehling, Roland Fried, Isabel García, and Martin Wendler. Change-point detection under dependence based on two-sample U-statistics. In *Asymptotic Laws and Methods in Stochastics*, volume 76 of *Fields Institute Communications*, pages 195–220. Springer, New York, 2015.

[14] Herold Dehling, Kata Vuk, and Martin Wendler. Change-point detection based on weighted two-sample U-statistics. *Electronic Journal of Statistics*, 16(1):862–891, 2022.

[15] Alexander Dürre, David E. Tyler, and Daniel Vogel. On the eigenvalues of the spatial sign covariance matrix in more than two dimensions. *Statistics & Probability Letters*, 111:80–85, 2016.

[16] Dietmar Ferger. A functional law of the iterated logarithm for U-statistic type processes. *Acta Applicandae Mathematicae*, 78:115–120, 2003.

[17] Evarist Giné, Stanisław Kwapień, Rafał Latała, and Joel Zinn. The LIL for canonical U-statistics of order 2. *The Annals of Probability*, 29(1):520–557, 2001.

[18] Felix Gnettner and Claudia Kirch. A new and flexible class of sharp asymptotic time-uniform confidence sequences. *Statistics & Probability Letters*, 226:110462, 2025.

[19] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[20] Peter Hall. On the invariance principle for U-statistics. *Stochastic Processes and their Applications*, 9(2):163–174, 1979.

[21] Fang Han and Han Liu. ECA: High-dimensional elliptical component analysis in non-gaussian distributions. *Journal of the American Statistical Association*, 113(521):252–268, 2018.

[22] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

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

[24] Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[25] Claudia Kirch and Christina Stoehr. Sequential change point tests based on U-statistics. *Scandinavian Journal of Statistics*, 49(3):1184–1214, 2022.

[26] Vladimir Koltchinskii and Evarist Giné. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167, 2000.

[27] Vladimir S. Korolyuk and Yuri V. Borovskich. *Theory of U-Statistics*. Kluwer Academic Publishers, Dordrecht, 1994.

[28] Alan J. Lee. *U-Statistics: Theory and Practice*. Marcel Dekker, New York, 1990.

[29] Weijia Li, Leheng Cai, and Qirui Hu. Strong gaussian approximation for u-statistics in high dimensions and beyond. *arXiv preprint arXiv:2603.10595*, 2026.

[30] Henry B. Mann and Donald R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.

[31] Nitis Mukhopadhyay and Inger Vik. Asymptotic results for stopping times based on U-statistics. *Sequential Analysis*, 4(1–2):83–109, 1985.

[32] Nitis Mukhopadhyay and Inger Vik. Convergence rates for two-stage confidence intervals based on U-statistics. *Annals of the Institute of Statistical Mathematics*, 40(1):111–117, 1988.

[33] Masoud M. Nasari. Studentized processes of U-statistics, 2009.

[34] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

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

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

[37] Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11:905–934, 2010.

[38] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

[39] Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York, 1980.

[40] Raymond N. Sproule. Sequential nonparametric fixed-width confidence intervals for U-statistics. *The Annals of Statistics*, 13(1):228–235, 1985.

[41] Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

[42] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

[43] Jean Ville. *Étude Critique de la Notion de Collectif*. Gauthier-Villars, Paris, 1939.

[44] Ian Waudby-Smith, David Arbour, Ritwik Sinha, Edward H. Kennedy, and Aaditya Ramdas. Time-uniform central limit theory and asymptotic confidence sequences. *The Annals of Statistics*, 52(6):2613–2640, 2024.

[45] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 02 2024.

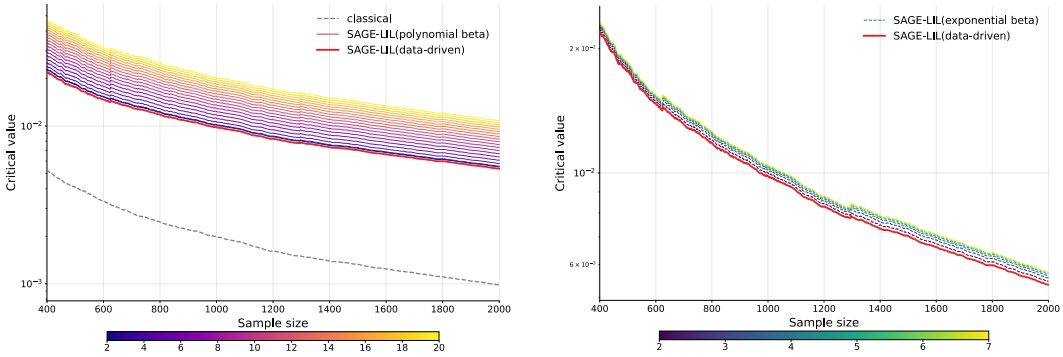
[46] Wojciech Zaremba, Arthur Gretton, and Matthew B. Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *Advances in Neural Information Processing Systems 26*, pages 755–763, 2013.

A Additional numerical results

A.1 Results of sensitivity experiments

Figure A.1 examines the sensitivity of the SAGE-LIL boundary to the spectral weight allocation, as discussed in Section 5. We compare the dynamically updated data-driven allocation $\widehat{\beta}_{\ell,n} \propto (\widehat{\lambda}_{\ell,n})_+$ with two families of deterministic allocations: polynomial weights $\beta_\ell \propto \ell^{-b}$ and exponential weights $\beta_\ell \propto e^{-c(\ell-1)}$.

The boundary is sensitive to the choice of deterministic weights. In particular, rapidly decaying allocations place too little weight on higher spectral coordinates and can lead to substantially wider boundaries, as seen for large b in panel (a). Across the range of allocations considered in both panels, the data-driven choice yields the tightest boundary. These numerical results are consistent with the theoretical motivation for the adaptive allocation in Proposition C.3.



(a) Polynomial decay ($\beta_\ell \propto \ell^{-b}$)

(b) Exponential decay ($\beta_\ell \propto e^{-c(\ell-1)}$)

Figure A.1: Sensitivity of the SAGE-LIL boundary to the spectral weight allocation β_ℓ. The colorbars indicate the decay-rate parameters: the polynomial degree $b \in [2, 20]$ in panel (a), and the exponential rate $c \in [2, 7]$ in panel (b). The data-driven allocation is shown separately in each panel.

A.2 AsympCSs for GMD under Laplace and t_{10} distributions

Figures A.2-A.3 present the performance of the proposed AsympCSs for GMD under Laplace and t_{10} distributions. In our implementation, the data dimension is set to $d = 1$ for the GMD, sample variance, and MMD tests, and $d = 2$ for the Spatial Kendall's tau.

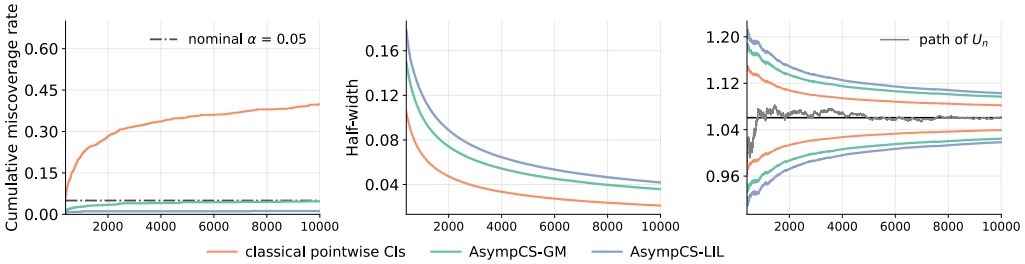


Figure A.2: Left: cumulative miscoverage rates of the proposed AsympCSs and the classical pointwise CIs for GMD under the Laplace distribution. Middle: averaged half-widths of the proposed AsympCSs and the classical pointwise CIs. Right: a single sample path of the statistics U_n for GMD alongside the three boundaries, where the black horizontal line indicates the true parameter. The horizontal axis in all three panels represents the sample size.

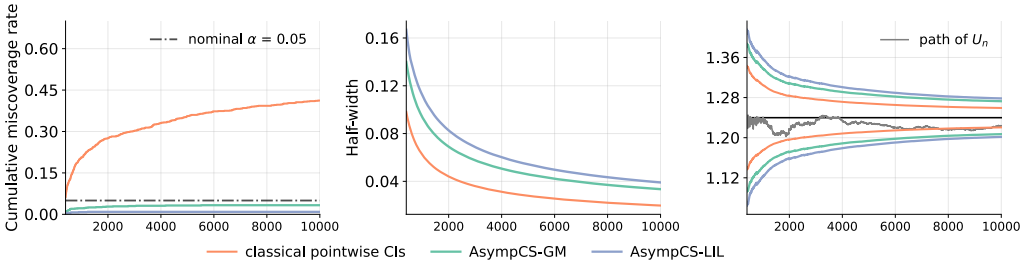


Figure A.3: Left: cumulative miscoverage rates of the proposed AsympCSs and the classical pointwise CIs for GMD under the t_{10} distribution. Middle: averaged half-widths of the proposed AsympCSs and the classical pointwise CIs. Right: a single sample path of the statistics U_n for GMD alongside the three boundaries, where the black horizontal line indicates the true parameter. The horizontal axis in all three panels represents the sample size.

A.3 Two-sample sequential testing based on MMD under Laplace and t_{10} null distributions

Figures A.4-A.5 present the performance of the sequential testing procedures using the proposed SAGE boundaries for sequential two-sample test with MMD kernel statistics under Laplace and t_{10} distributions.

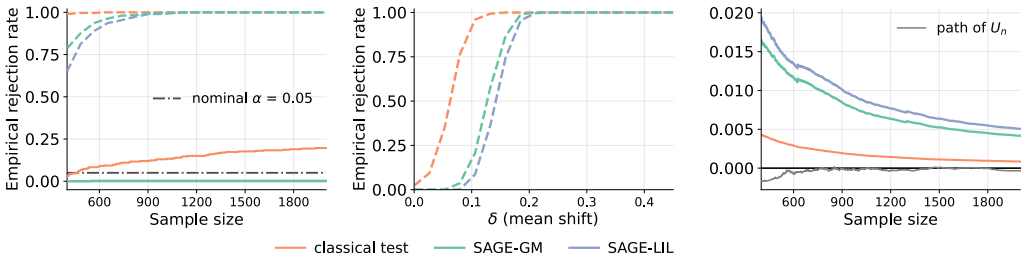


Figure A.4: Left: the size and power comparison of the proposed sequential testing procedure using the SAGE boundaries and the classical test procedure for sequential two-sample test with MMD kernel statistics under Laplace distribution. Dashed lines represent the power under H_1 with $\delta = 0.3$, while solid lines represent the size under H_0 with $\delta = 0$. Middle: the empirical power over mean shifts $\delta \in [0, 0.45]$ at sample size $n = 2000$. Right: a single sample path of the MMD test statistic U_n alongside the three boundaries under H_0.

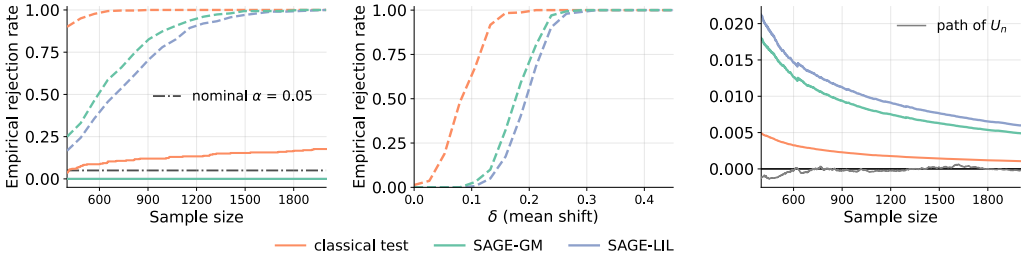


Figure A.5: Left: the size and power comparison of the proposed sequential testing procedure using the SAGE boundaries and the classical test procedure for sequential two-sample test with MMD kernel statistics under t_{10} distribution. Dashed lines represent the power under H_1 with $\delta = 0.3$, while solid lines represent the size under H_0 with $\delta = 0$. Middle: the empirical power over mean shifts $\delta \in [0, 0.45]$ at sample size $n = 2000$. Right: a single sample path of the MMD test statistic U_n alongside the three boundaries under H_0.

A.4 Additional experiments

Figures A.6-A.8 present the performance of the proposed AsympCSs for sample variance under the same three distributions as in Section 5.

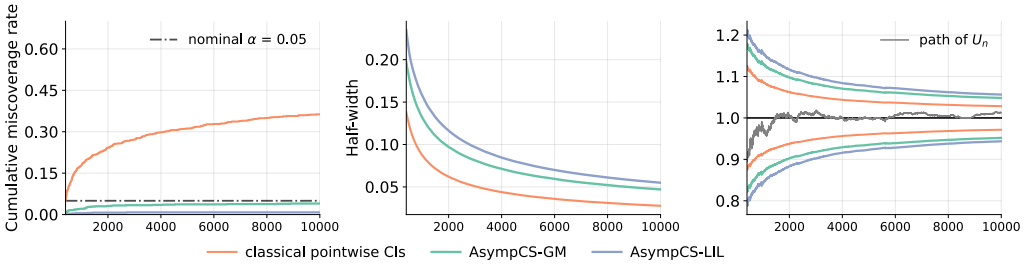


Figure A.6: Left: cumulative miscoverage rates of the proposed AsympCSs and the classical pointwise CIs for sample variance under the Gaussian distribution. Middle: averaged half-widths of the proposed AsympCSs and the classical pointwise CIs. Right: a single sample path of the statistics U_n for sample variance alongside the three boundaries, where the black horizontal line indicates the true parameter. The horizontal axis in all three panels represents the sample size.

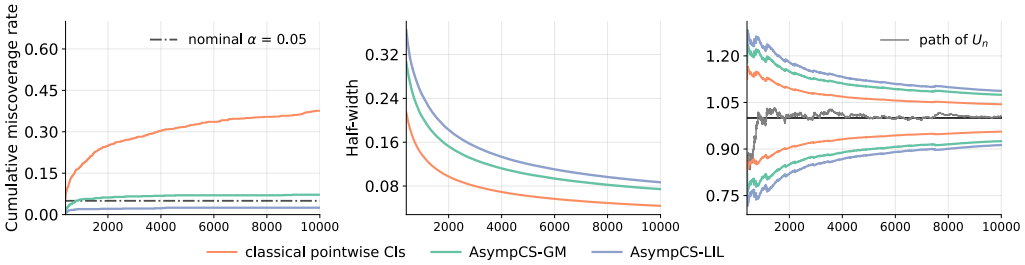


Figure A.7: Left: cumulative miscoverage rates of the proposed AsympCSs and the classical pointwise CIs for sample variance under the Laplace distribution. Middle: averaged half-widths of the proposed AsympCSs and the classical pointwise CIs. Right: a single sample path of the statistics U_n for sample variance alongside the three boundaries, where the black horizontal line indicates the true parameter. The horizontal axis in all three panels represents the sample size.

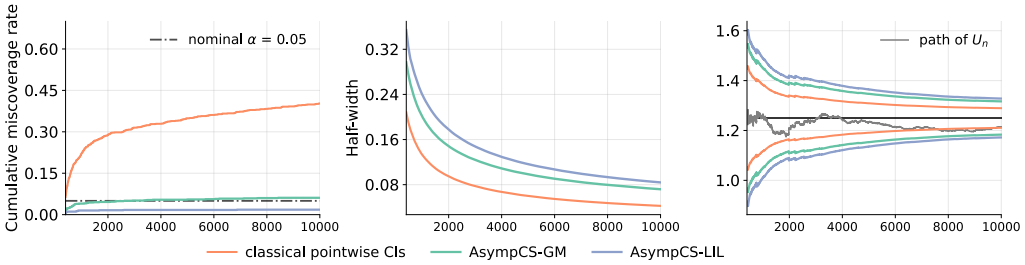


Figure A.8: Left: cumulative miscoverage rates of the proposed AsympCSs and the classical pointwise CIs for sample variance under the t_{10} distribution. Middle: averaged half-widths of the proposed AsympCSs and the classical pointwise CIs. Right: a single sample path of the statistics U_n for sample variance alongside the three boundaries, where the black horizontal line indicates the true parameter. The horizontal axis in all three panels represents the sample size.

We notice that in Figures A.7-A.8, the cumulative miscoverage rate of AsympCS-GM slightly exceeds the nominal level. Theoretically, the Gaussian approximation errors for U_n and the consistency of the variance estimator $\widehat{\sigma}_n^2$ depend on higher-order moments of the kernel. For the sample variance, such moment dependence is more pronounced, so heavier-tailed distributions require larger sample sizes

for the asymptotic approximation to become accurate. Increasing the cold-start parameter m mitigates this finite-sample under-coverage by delaying monitoring until sufficient data have accumulated.

Figures A.9-A.11 present the performance of the proposed AsympCSs estimation for Spatial Kendall's tau under the three distributions. Here, we detail the data generation process for the Spatial Kendall's tau estimation. Unlike the other three kernels we adopted, which utilize one-dimensional data drawn from the three specified distributions, Spatial Kendall's tau requires bivariate observations. To accommodate this, we generate two-dimensional data vectors $\boldsymbol{X}_i = (X_{i1}, X_{i2})^\top \in \mathbb{R}^2$ with a correlation structure. Specifically, the dependence is governed by a shape matrix $\boldsymbol{\Sigma}$, defined as:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Throughout our simulations, the shape correlation is set to $\rho = 0.6$. For the two-dimensional elliptical distributions, $\boldsymbol{X}_i$ has the scale mixture of normals representation:

$$\boldsymbol{X}_i = \sqrt{W_i}\mathbf{A}\boldsymbol{Z}_i, \tag{14}$$

where $\boldsymbol{Z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ is a standard two-dimensional Gaussian vector, and $\mathbf{A}$ is the Cholesky factor satisfying $\mathbf{A}\mathbf{A}^\top = \boldsymbol{\Sigma}$. The multiplier $W_i > 0$ is a scalar random variable, independent of $\boldsymbol{Z}_i$, that dictates the tail heaviness of the resulting bivariate distribution.

In our simulations, we use three different W_i:

- **Gaussian**: $W_i = 1$ a.s., yielding the bivariate normal distribution with correlation ρ.
- **Student's** t_{10}: $W_i = \nu/V_i$, where $V_i \sim \chi^2_\nu$ with degrees of freedom $\nu = 10$.
- **Laplace**: $W_i \sim \mathrm{Exp}(1)$, which generates the bivariate symmetric Laplace distribution.

By this representation, we first sample the standard two-dimensional Gaussian data $\boldsymbol{Z}_i$ and obtain $\boldsymbol{X}_i$ from (14), then compute the corresponding U-statistics. By mathematical derivation,

$$\theta = \mathbb{E}\left\{\frac{(X_{11} - X_{21})(X_{12} - X_{22})}{\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_2^2}\mathbf{1}\{\boldsymbol{X}_1 \neq \boldsymbol{X}_2\}\right\} = \left(1 - \sqrt{1-\rho^2}\right)/(2\rho).$$

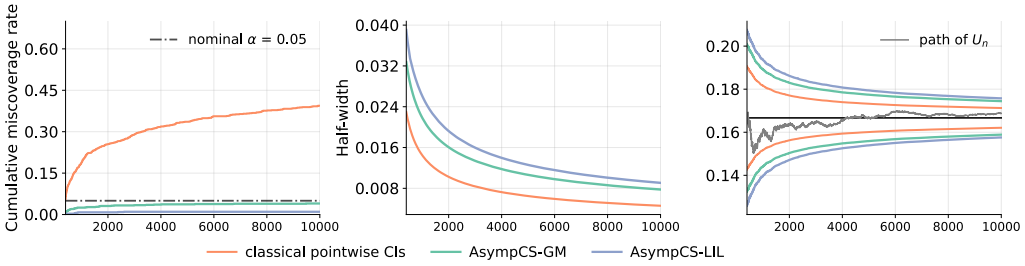


Figure A.9: Left: cumulative miscoverage rates of the proposed AsympCSs and the classical pointwise CIs for Spatial Kendall's tau under the Gaussian distribution. Middle: averaged half-widths of the proposed AsympCSs and the classical pointwise CIs. Right: a single sample path of the statistics U_n for Spatial Kendall's tau alongside the three boundaries, where the black horizontal line indicates the true parameter. The horizontal axis in all three panels represents the sample size.

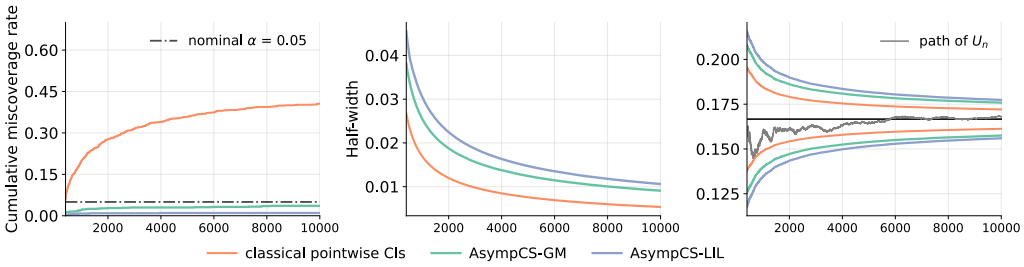


Figure A.10: Left: cumulative miscoverage rates of the proposed AsympCSs and the classical pointwise CIs for Spatial Kendall's tau under the Laplace distribution. Middle: averaged half-widths of the proposed AsympCSs and the classical pointwise CIs. Right: a single sample path of the statistics U_n for Spatial Kendall's tau alongside the three boundaries, where the black horizontal line indicates the true parameter. The horizontal axis in all three panels represents the sample size.

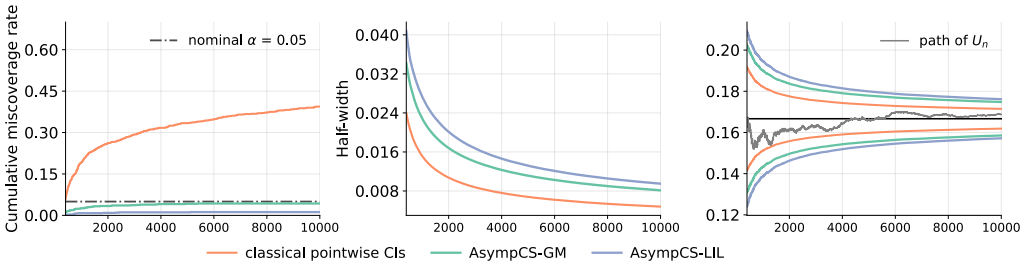


Figure A.11: Left: cumulative miscoverage rates of the proposed AsympCSs and the classical pointwise CIs for Spatial Kendall's tau under the t_{10} distribution. Middle: averaged half-widths of the proposed AsympCSs and the classical pointwise CIs. Right: a single sample path of the statistics U_n for Spatial Kendall's tau alongside the three boundaries, where the black horizontal line indicates the true parameter. The horizontal axis in all three panels represents the sample size.

A.5 Acceleration of Algorithm 2

A direct implementation of Algorithm 2 requires recomputing the spectrum of an $n \times n$ centered Gram matrix at every monitoring time. Although the U-statistic itself admits an exact $\mathcal{O}(n)$ recursive update at time n, a full eigendecomposition costs $\mathcal{O}(n^3)$, leading to a cumulative spectral cost of

$$\sum_{n \le T} \mathcal{O}(n^3) = \mathcal{O}(T^4)$$

up to horizon T. We therefore accelerate Algorithm 2, while continuing to update the U-statistic and the trace estimator using all available observations.

Geometric refresh and subsampling. Let $t_0 = m$ and $t_{r+1} = \lceil \rho t_r \rceil$ where $\rho > 1$ be a geometric sequence of refresh times. At time t_r, select a data-independent subset $\mathcal{I}_r \subset \{1, \dots, t_r\}$ whose cardinality is $|\mathcal{I}_r| = N_r \asymp t_r^{\varpi}$ where $0 < \varpi < 1$.

The subset may be deterministic or randomly selected independently of the observed data values. Using the observations indexed by $\mathcal{I}_r$, form the centered sub-Gram matrix

$$\widehat{K}_r^{\mathrm{sub}}(i,j) = h(X_i, X_j) - U_{t_r}, \qquad i, j \in \mathcal{I}_r,$$

and let $\{\widehat{\lambda}_{\ell,r}^{\mathrm{sub}}\}_{\ell \ge 1}$ denote the eigenvalues of $N_r^{-1}\widehat{K}_r^{\mathrm{sub}}$, under the same signed indexing convention as the population spectrum. Retaining $L_r \asymp t_r^a$ eigenvalues, define

$$\begin{aligned}
\widehat{\Lambda}_r^+ &= \sum_{\ell \le L_r} \big(\widehat{\lambda}_{\ell,r}^{\mathrm{sub}}\big)_+, \\
\widehat{\Lambda}_{\beta,r}^+ &= \sum_{\ell \le L_r} \log(1/\beta_\ell)\big(\widehat{\lambda}_{\ell,r}^{\mathrm{sub}}\big)_+, \\
\widehat{\Lambda}_{\beta,g,r}^+ &= \sum_{\ell \le L_r} \{g^{-1}(\alpha\beta_\ell)\}^2\big(\widehat{\lambda}_{\ell,r}^{\mathrm{sub}}\big)_+.
\end{aligned}$$

These spectral quantities are held fixed for $t_r \le n < t_{r+1}$. And the trace functional requires no eigendecomposition and is updated separately at every monitoring time:

$$\widehat{\Lambda}_n = \frac{1}{n}\sum_{i=1}^{n} h(X_i, X_i) - U_n.$$

Thus, only the spectral summaries are refreshed geometrically, whereas U_n and $\widehat{\Lambda}_n$ continue to use the full data stream.

The following proposition shows that this modification preserves the spectral approximation condition required by Assumption 1.

Proposition A.1. *At each refresh time t_r, suppose $\sup_{\ell \ge 1}\left|\widehat{\lambda}_{\ell,r}^{\mathrm{sub}} - \lambda_\ell\right| = \mathcal{O}_{\mathrm{a.s.}}(d_r)$ for some $d_r \downarrow 0$. Let $L_r \uparrow \infty$ and, for some $0 < \iota < 1/2$, assume*

$$d_r \sum_{\ell \le L_r}\{1 + \log(1/\beta_\ell)\} + \sum_{\ell > L_r}(\lambda_\ell)_+\{1 + \log(1/\beta_\ell)\} = o(t_r^{-\iota}),$$

together with $|\widehat{\Lambda}_n - \Lambda| = o_{\mathrm{a.s.}}(n^{-\iota})$. If the refreshed spectral estimates are held fixed over $t_r \le n < t_{r+1}$, then, uniformly on each refresh interval,

$$|\widehat{\Lambda}_n - \Lambda| + |\widehat{\Lambda}_n^+ - \Lambda^+| + |\widehat{\Lambda}_{\beta,n}^+ - \Lambda_\beta^+| + |\widehat{\Lambda}_{\beta,g,n}^+ - \Lambda_{\beta,g}^+| = o_{\mathrm{a.s.}}(n^{-\iota}).$$

Hence Assumption 1 remains valid under geometric spectral refreshing and subsampling. In particular, under the conditions of Theorems 4 and 5, the conclusion of Theorem 6 continues to hold.

Proof. Let $\delta_r = \sup_{\ell \ge 1}|\widehat{\lambda}_{\ell,r}^{\mathrm{sub}} - \lambda_\ell|$. Since $x \mapsto x_+$ is 1-Lipschitz and, for fixed $\alpha \in (0,1)$, $\{g^{-1}(\alpha\beta_\ell)\}^2 \lesssim 1 + \log(1/\beta_\ell)$, all three spectral errors are bounded simultaneously by

$$\begin{aligned}
&|\widehat{\Lambda}_r^+ - \Lambda^+| + |\widehat{\Lambda}_{\beta,r}^+ - \Lambda_\beta^+| + |\widehat{\Lambda}_{\beta,g,r}^+ - \Lambda_{\beta,g}^+| \\
&\qquad \lesssim \delta_r \sum_{\ell \le L_r}\{1 + \log(1/\beta_\ell)\} + \sum_{\ell > L_r}(\lambda_\ell)_+\{1 + \log(1/\beta_\ell)\}.
\end{aligned}$$

The assumed rate and $\delta_r = \mathcal{O}_{\mathrm{a.s.}}(d_r)$ therefore give an $o_{\mathrm{a.s.}}(t_r^{-\iota})$ bound at every refresh time. Because $t_{r+1} = \lceil \rho t_r \rceil$, we have $t_r \asymp n$ uniformly for $t_r \le n < t_{r+1}$, so the same bound is $o_{\mathrm{a.s.}}(n^{-\iota})$ throughout each refresh interval. Combining this with the assumed rate for $\widehat{\Lambda}_n$ proves the claim. □

A concrete rate choice and complexity. Under the same empirical-operator regularity conditions as in Proposition C.2, the subsampled empirical operator based on N_r observations satisfies, by Weyl's inequality,

$$\sup_{\ell \ge 1} \left| \widehat{\lambda}_{\ell,r}^{\mathrm{sub}} - \lambda_\ell \right| = \mathcal{O}_{\mathrm{a.s.}} \left\{ \left(\frac{\log t_r}{N_r} \right)^{\kappa} \right\}, \qquad 0 < \kappa \le \frac{1}{2},$$

where we use $N_r \asymp t_r^{\varpi}$ and hence $\log N_r \asymp \log t_r$. The additional centering by U_{t_r} instead of θ is a rank-one perturbation of operator norm $|U_{t_r} - \theta|$ and is absorbed into the same rate. If $(\lambda_\ell)_+ \lesssim \ell^{-q}$ for some $q > 1$, and $L_r \asymp t_r^a$ and $\beta_\ell \asymp \ell^{-b}$ with $b > 1$, then Proposition A.1 applies for every

$$0 < \iota < \min\{\varpi\kappa - a,\ a(q-1)\}, \qquad 0 < a < \varpi\kappa.$$

In particular, under the standard empirical-operator rate $\kappa = 1/2$, choosing $\varpi = 2/3$ permits any $a < 1/3$; we use $a = 1/4$ in the implementation below.

Because the refresh times grow geometrically, only $\mathcal{O}(\log T)$ spectral decompositions are required up to horizon T. Since $N_r \asymp t_r^{\varpi}$, their cumulative cost is $\mathcal{O}(T^{3\varpi})$, while constructing the corresponding sub-Gram matrices costs $\mathcal{O}(T^{2\varpi})$. With $\varpi = 2/3$, the cumulative spectral cost is therefore $\mathcal{O}(T^2)$, matching the $\mathcal{O}(T^2)$ cost of the exact pairwise updates for U_n. Hence the overall computational complexity is reduced from $\mathcal{O}(T^4)$ to $\mathcal{O}(T^2)$, while the storage required for the latest Gram matrix decreases from $\mathcal{O}(T^2)$ to $\mathcal{O}(T^{4/3})$.

Numerical comparison. We compared the accelerated implementation with the original full-spectrum implementation under the same Monte Carlo design used for the degenerate MMD experiments. We used $\beta_\ell = \ell^{-2}/\zeta(2), \varpi = 2/3, a = 1/4$ together with the geometric refresh schedule above. The accelerated procedure produced critical-value trajectories and empirical sequential coverage nearly indistinguishable from those of the full-spectrum implementation. Meanwhile, the total runtime decreased from approximately $9{,}216$ seconds to 12.89 seconds, corresponding to a roughly 715-fold speedup. Thus, in the settings considered here, geometric spectral refreshing substantially reduces the computational cost without materially changing the finite-sample behavior of the SAGE procedure.

B Technical proofs of theorems

B.1 Proof of Theorem 1

Since $\mathbb{E}h^2(X_1, X_2) < \infty$, applying Theorem 5.3.3 in [39], we obtain that for any $\epsilon > 0$,

$$|R_n| = o_{a.s.}\left(\frac{(\log n)^{1/2+\epsilon}}{n} \right).$$

Denote $\sigma^2 = \mathrm{Var}(h_1(X))$. Applying Theorem 2.6.3 of [8], there exists a Wiener process $\mathcal{W}(\cdot)$ such that,

$$\left| \frac{1}{n} \sum_{i=1}^{n} h_1(X_i) - \frac{\sigma}{n} \mathcal{W}(n) \right| = o_{a.s.}\left(n^{-1+1/(2+\delta)} \right).$$

Therefore, we derive that

$$\left| U_n - \theta - \frac{2\sigma}{n} \mathcal{W}(n) \right| = o_{a.s.}\left(n^{-1+1/(2+\delta)} \right) = o_{a.s.}\left(\sqrt{\frac{\log\log n}{n}} \right).$$

By defining the i.i.d. standard Gaussian variables as the exact independent increments of the Wiener process on this enriched space, $Z_i := \mathcal{W}(i) - \mathcal{W}(i-1)$, we obtain the strict pathwise equality:

$$\frac{2\sigma}{n} \mathcal{W}(n) = \frac{2\sigma}{n} \sum_{i=1}^{n} Z_i.$$

B.2 Proof of Theorem 2

Define $\widehat{h}_{1,n}(X_i) = A_{i,n} - U_n$, where

$$A_{i,n} := \frac{1}{n-1}\sum_{j\neq i} h(X_i, X_j), \qquad 1 \leq i \leq n.$$

Thus,

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \widehat{h}_{1,n}^2(X_i) = \frac{1}{n}\sum_{i=1}^n (A_{i,n} - U_n)^2 = \frac{1}{n}\sum_{i=1}^n A_{i,n}^2 - U_n^2.$$

Note that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n A_{i,n}^2 &= \frac{1}{n(n-1)^2}\sum_{i=1}^n \sum_{j\neq i}\sum_{k\neq i} h(X_i, X_j)h(X_i, X_k) \\ &= \frac{1}{n(n-1)^2}\sum_{i\neq j\neq k} h(X_i, X_j)h(X_i, X_k) \\ &+ \frac{1}{n(n-1)^2}\sum_{i\neq j} h^2(X_i, X_j)\end{aligned}$$

Let

$$K(x, y, z) := h(x, y)h(x, z).$$

Since $\mathbb{E}[h(X_1, X_2)^4] < \infty$, Cauchy-Schwarz inequality yields

$$\mathbb{E}|K(X_1, X_2, X_3)|^2 \leq \left(\mathbb{E}[h(X_1, X_2)^4]\right)^{1/2}\left(\mathbb{E}[h(X_1, X_3)^4]\right)^{1/2} < \infty.$$

Thus, Theorem C in Section 5.4 of [39] gives

$$\left|\frac{1}{n(n-1)^2}\sum_{i\neq j\neq k} K(X_i, X_j, X_k) - \mathbb{E}K(X_1, X_2, X_3)\right| = \mathcal{O}_{a.s.}\left(\sqrt{\frac{\log\log n}{n}}\right).$$

Similarly, one obtains

$$\left|\frac{1}{n(n-1)^2}\sum_{i\neq j} h^2(X_i, X_j)\right| = \mathcal{O}_{a.s.}\left(\frac{1}{n}\right).$$

Further Observe that

$$\mathbb{E}[K(X_1, X_2, X_3)] = \mathbb{E}\left[\mathbb{E}[h(X_1, X_2) \mid X_1]\mathbb{E}[h(X_1, X_3) \mid X_1]\right].$$

Hence,

$$\left|\frac{1}{n}\sum_{i=1}^n A_{i,n}^2 - \mathbb{E}\left[\mathbb{E}[h(X_1, X_2) \mid X_1]\mathbb{E}[h(X_1, X_3) \mid X_1]\right]\right| = \mathcal{O}_{a.s.}\left(\sqrt{\frac{\log\log n}{n}}\right). \tag{15}$$

Since $\mathbb{E}h^2(X_1, X_2) < \infty$, Theorem C in Section 5.4 of [39] gives

$$|U_n - \theta| = \mathcal{O}_{a.s.}\left(\sqrt{\frac{\log\log n}{n}}\right),$$

which entails that

$$|U_n^2 - \theta^2| \leq |U_n - \theta|\left(2|\theta| + o_{a.s.}(1)\right) = \mathcal{O}_{a.s.}\left(\sqrt{\frac{\log\log n}{n}}\right). \tag{16}$$

Therefore, for any $\epsilon > 0$, using (6), (15), and (16),

$$\left|\widehat{\sigma}_n^2 - \sigma^2\right| = \mathcal{O}_{a.s.}\left(\sqrt{\frac{\log\log n}{n}}\right) = o_{a.s.}\left(n^{-1/2+\epsilon}\right).$$

B.3 Proof of Theorem 3

The result directly holds by Theorems 1-2 and Theorems 2.4, 2.8 of [44].

B.4 Proof of Theorem 4

For degenerate kernel function $h(x, y)$, it can be represented by

$$h\left(X_i, X_j\right) - \theta = \sum_{k=1}^{\infty} \lambda_k \psi_k\left(X_i\right) \psi_k\left(X_j\right),$$

where $|\lambda_k| \geqslant |\lambda_{k+1}|$, $\mathbb{E}\psi_k(X_i) = 0$, $\mathbb{E}\psi_k^2\left(X_i\right) = 1$, $\mathbb{E}\left\{\psi_k\left(X_i\right) \psi_{k'}\left(X_i\right)\right\} = 0$ for $k \neq k'$, and

$$\mathbb{E}\left\{h\left(X_i, x\right) \psi_k\left(X_i\right)\right\} = \lambda_k \psi_k(x).$$

Let $F(\cdot)$ be the distribution function of X. Theorem 2, together with the diagonal decomposition and spectral representation in Section 3 of [12] gives independent standard Brownian bridges $\mathcal{B}_i$ on $[0, 1]$ on an enriched probability space such that

$$\sum_{i,j=1}^{n} \{h(X_i, X_j) - \theta\} = n\{\mathbb{E}h(X, X) - \theta\} + \sum_{k=1}^{\infty} \lambda_k \left[\left\{ \sum_{i=1}^{n} \int_0^1 \psi_k(F^{-1}(x))\, d\mathcal{B}_i(x) \right\}^2 - n \right] + \mathcal{o}_{a.s.}(n \log\log n).$$

The centered spectral series converges in L^2 for each n, since $\sum_k \lambda_k^2 < \infty$; its tail after coordinate r has variance $2n^2 \sum_{k>r} \lambda_k^2$. Besides, the law of the iterated logarithm gives

$$\sum_{i=1}^{n} \{h(X_i, X_i) - \mathbb{E}h(X_i, X_i)\} = \mathcal{O}_{a.s.}\left(\sqrt{n \log\log n}\right).$$

Subtracting the diagonal terms from the preceding double-sum approximation, and using

$$n(n-1)(U_n - \theta) = \sum_{i,j \leq n} \{h(X_i, X_j) - \theta\} - \sum_{i=1}^{n} \{h(X_i, X_i) - \theta\},$$

we obtain

$$n(n-1)(U_n - \theta) = \sum_{k=1}^{\infty} \lambda_k \left[\left\{ \sum_{i=1}^{n} \int_0^1 \psi_k(F^{-1}(x))\, d\mathcal{B}_i(x) \right\}^2 - n \right] + \mathcal{o}_{a.s.}(n \log\log n).$$

The law of the iterated logarithm for canonical degree-two U-statistics also gives $U_n - \theta = \mathcal{O}_{a.s.}(\log\log n/n)$. Consequently,

$$(U_n - \theta) - \frac{n-1}{n}(U_n - \theta) = \frac{U_n - \theta}{n} = \mathcal{O}_{a.s.}\left(\frac{\log\log n}{n^2}\right).$$

Dividing the preceding numerator approximation by n^2 therefore yields

$$\left| U_n - \theta - \sum_{k=1}^{\infty} \lambda_k \left[\left\{ \frac{1}{n} \sum_{i=1}^{n} \int_0^1 \psi_k(F^{-1}(x))\, d\mathcal{B}_i(x) \right\}^2 - \frac{1}{n} \right] \right| = \mathcal{o}_{a.s.}\left(\frac{\log\log n}{n}\right).$$

We explicitly define the random variables Z_{ik} on this enriched probability space as the stochastic integrals:

$$Z_{ik} := \int_0^1 \psi_k(F^{-1}(x)) d\mathcal{B}_i(x).$$

Then we have

$$\sum_{k=1}^{\infty} \lambda_k \left[\left\{ \frac{1}{n} \sum_{i=1}^{n} \int \psi_k(F^{-1}(x)) d\mathcal{B}_i(x) \right\}^2 - \frac{1}{n} \right] = \sum_{k=1}^{\infty} \lambda_k \left\{ \left(\frac{1}{n} \sum_{i=1}^{n} Z_{ik} \right)^2 - \frac{1}{n} \right\}.$$

Using $\mathbb{E}\psi_k(X) = 0$, $\mathbb{E}\psi_k^2(X) = 1$, $\mathbb{E}\psi_k(X)\psi_{k'}(X) = 0$ for $k \neq k'$, it follows that $\{Z_{ik}\}$ are exactly mutually independent standard Gaussian variables. On a further extension, these can be taken as the unit-time increments of independent standard Brownian motions W_k, so that $W_k(n) = \sum_{i=1}^{n} Z_{ik}$, completing the proof.

B.5 Proof of Theorem 5

For each $\ell \geq 1$, set $\alpha_\ell := \alpha\beta_\ell$ and define $E_\ell := \{\forall n \geq m : \ |W_\ell(n)|/n < \gamma_{\alpha_\ell,m}(n)\}$. By Definition 3, $\mathbb{P}(E_\ell^c) \leq \alpha_\ell$. Let $E := \bigcap_{\ell\geq 1} E_\ell$. Then

$$\mathbb{P}(E^c) \leq \sum_{\ell\geq 1} \mathbb{P}(E_\ell^c) \leq \sum_{\ell\geq 1} \alpha_\ell = \alpha \sum_{\ell\geq 1} \beta_\ell \leq \alpha.$$

On E,

$$\begin{aligned}
\frac{1}{n^2}\sum_{\ell\geq 1} \lambda_\ell \left\{W_\ell^2(n) - n\right\} &= \frac{1}{n^2}\sum_{\ell:\lambda_\ell>0} \lambda_\ell \left\{W_\ell^2(n) - n\right\} + \frac{1}{n^2}\sum_{\ell:\lambda_\ell<0} \lambda_\ell \left\{W_\ell^2(n) - n\right\} \\
&\leq \frac{1}{n^2}\sum_{\ell:\lambda_l>0} \lambda_\ell W_\ell^2(n) - \frac{1}{n}\sum_{\ell\geq 1}\lambda_\ell \\
&\leq \sum_{\ell:\lambda_l>0} \lambda_\ell \gamma_{\alpha_\ell,m}^2(n) - \frac{1}{n}\sum_{\ell\geq 1}\lambda_\ell.
\end{aligned}$$

Similarly,

$$\frac{1}{n^2}\sum_{\ell\geq 1} \lambda_\ell \left\{W_\ell^2(n) - n\right\} \geq \sum_{\ell:\lambda_l<0} \lambda_\ell \gamma_{\alpha_\ell,m}^2(n) - \frac{1}{n}\sum_{\ell\geq 1}\lambda_\ell.$$

For $\Upsilon^{\mathbf{LIL}}_{\alpha,m}(n)$:

$$\begin{aligned}
&\sum_{\ell:\lambda_l>0} \lambda_\ell \left\{\gamma^{\mathrm{LIL}}_{\alpha_\ell,m}(n)\right\}^2 - \frac{1}{n}\sum_{\ell\geq 1}\lambda_\ell \\
=&\frac{\left(\eta^{1/4}+\eta^{-1/4}\right)^2}{2n}\left[\left\{s\log\log\left(\max\left\{\frac{\eta n}{m}, e\right\}\right) + \log\frac{\zeta(s)}{\alpha(\log\eta)^s}\right\}\Lambda^+ + \Lambda^+_\beta\right] - \frac{\Lambda}{n}.
\end{aligned}$$

For $\Upsilon^{\mathbf{GM}}_{\alpha,m}(n)$: Let

$$g(a) = 2\{1 - \Phi(a) + a\phi(a)\}, \qquad a \geq 0.$$

We first study the tail order of g. Since

$$g'(a) = 2\{-\phi(a) + \phi(a) - a^2\phi(a)\} = -2a^2\phi(a) < 0, \qquad a > 0,$$

g is strictly decreasing on $[0,\infty)$, and

$$g(0) = 1, \qquad g(a) \downarrow 0 \quad \text{as } a \to \infty.$$

Hence $g^{-1}(\cdot)$ is well defined on $(0,1]$. By Mills' ratio, for $a > 0$,

$$\frac{a}{1+a^2}\phi(a) \leq 1 - \Phi(a) \leq \frac{\phi(a)}{a}.$$

Therefore, for $a \geq 1$,

$$2a\phi(a) \leq g(a) = 2\{1 - \Phi(a) + a\phi(a)\} \leq 4a\phi(a),$$

which further yields that

$$g(a) \asymp a\phi(a) \asymp ae^{-a^2/2}, \qquad a \to \infty.$$

Consequently, as $x \downarrow 0$, $g^{-1}(x) \asymp \sqrt{\log(1/x)}$. Therefore, for any fixed $\alpha > 0$,

$$g^{-1}(\alpha\beta_\ell) \asymp \sqrt{\log\big(1/\beta_\ell\big)}.$$

Then, the condition $\sum_{\ell:\lambda_\ell>0} \lambda_\ell\{g^{-1}(\alpha\beta_\ell)\}^2 < \infty$ is asymptotically equivalent to

$$\sum_{\ell:\lambda_\ell>0} \lambda_\ell \log\big(1/\beta_\ell\big) < \infty.$$

Finally, it is clear that

$$\Upsilon^{\mathrm{GM}}_{\alpha,m}(n) = \frac{1}{n}\left[\Lambda^+\log\left(\frac{n}{m}\right) + \sum_{\ell:\lambda_\ell>0}\lambda_\ell\left\{g^{-1}(\alpha\beta_\ell)\right\}^2\right] - \frac{\Lambda}{n}.$$

B.6 Proof of Theorem 6

Theorems 4–5 provide an exact Gaussian-chaos CS, and (8) makes its strong-approximation error little-o of the SAGE scale. In addition, Assumption 1 makes the plug-in boundary numerically equivalent to its population version at the same scale. Therefore, the results in Theorems 2.4 and 2.8 of [44] establish the AsympCSs and the corresponding asymptotic time-uniform coverage property.

C Further discussion

C.1 Discussion on the lower boundaries

Recall that $\Lambda = \sum_{\ell \geq 1} \lambda_\ell$. Define $\Lambda^- = \sum_{\ell:\lambda_\ell<0} \lambda_\ell$, $\Lambda^-_\beta = \sum_{\ell:\lambda_\ell<0} \lambda_\ell \log(1/\beta_l)$ and $\Lambda^-_{\beta,g} = \sum_{\ell:\lambda_\ell<0} \lambda_\ell \left\{g^{-1}(\alpha\beta_\ell)\right\}^2$, where $g(a) = 2\{1-\Phi(a)+a\phi(a)\}$. We now introduce the lower boundary $\mathcal{L}_{\alpha,m}(\cdot)$. Suppose that $\sum_{\ell\geq 1}|\lambda_\ell| < \infty$, $\Lambda^-_\beta < \infty$ and $\{W_\ell(\cdot)\}_{\ell\geq 1}$ are independent standard Brownian motions. For any $\alpha \in (0,1)$, $m \geq 1$,

$$\mathbb{P}\left(\forall n \geq m:\ \frac{1}{n^2}\sum_{\ell\geq 1}\lambda_\ell\{W_\ell^2(n)-n\} \geq \mathcal{L}_{\alpha,m}(n)\right) \geq 1-\alpha,$$

where $\mathcal{L}_{\alpha,m}(n)$ can be chosen in either of the following two ways:

$$\mathcal{L}^{\mathrm{LIL}}_{\alpha,m}(n) = \frac{\left(\eta^{1/4}+\eta^{-1/4}\right)^2}{2n}\left[\left\{s\log\log\left(\max\left\{\frac{\eta n}{m}, e\right\}\right) + \log\frac{\zeta(s)}{\alpha(\log\eta)^s}\right\}\Lambda^- + \Lambda^-_\beta\right] - \frac{\Lambda}{n},$$

$$\mathcal{L}^{\mathrm{GM}}_{\alpha,m}(n) = \frac{1}{n}\left[\Lambda^-\log\left(\frac{n}{m}\right) + \Lambda^-_{\beta,g} - \Lambda\right].$$

If the kernel is positive semi-definite, then all eigenvalues are nonnegative, so the SAGE lower boundary degenerates to the deterministic term $\mathcal{L}^{\mathrm{LIL}}_{\alpha,m}(n) = \mathcal{L}^{\mathrm{GM}}_{\alpha,m}(n) = -\Lambda/n$.

Corollary 1 (Two-sided SAGE boundary)**.** *Let the deterministic weights satisfy $\beta_\ell > 0$ on every nonzero spectral coordinate and $\sum_\ell \beta_\ell \leq 1$. Under the conditions of Theorems 4 and 5, the oracle interval*

$$[U_n - \Upsilon_{\alpha,m}(n),\ U_n - \mathcal{L}_{\alpha,m}(n)]_{n\geq m}$$

is a two-sided $(1-\alpha)$-AsympCS. The same conclusion holds for empirical upper and lower boundaries if Assumption 1 and its negative-spectrum analogue hold for all quantities appearing in $\mathcal{L}_{\alpha,m}$.

Indeed, the single event $\bigcap_\ell\{|W_\ell(n)|/n \leq \gamma_{\alpha\beta_\ell,m}(n) \text{ for all } n \geq m\}$ has probability at least $1-\alpha$ and bounds the positive and negative spectral blocks simultaneously.

C.2 Discussion on Assumption 1

Proposition C.2 (Sufficient conditions for Assumption 1)**.** *Suppose that the conditions of Theorems 4-5 hold. Consider the centered Gram matrix*

$$\widehat{K}_n(i,j) = h(X_i,X_j) - U_n, \qquad 1 \leq i,j \leq n,$$

and let $(\widehat{\lambda}_{\ell,n})_{\ell\geq 1}$ be the eigenvalues of $n^{-1}\widehat{\mathbf{K}}_n$. Write $\widehat{\lambda}_\ell = \widehat{\lambda}_{\ell,n}$ when no confusion arises.

Set $L_n \asymp n^a$ and $\beta_\ell \asymp \ell^{-b}$ for some $b > 1$, and define $\widehat{\Lambda} = \frac{1}{n}\sum_{i=1}^n h(X_i,X_i) - U_n$, $\widehat{\Lambda}^+ = \sum_{\ell=1}^{L_n}(\widehat{\lambda}_\ell)_+$, $\widehat{\Lambda}^+_\beta = \sum_{\ell=1}^{L_n}(\widehat{\lambda}_\ell)_+\log(1/\beta_\ell)$, and $\widehat{\Lambda}^+_{\beta,g} = \sum_{\ell=1}^{L_n}(\widehat{\lambda}_\ell)_+\{g^{-1}(\alpha\beta_\ell)\}^2$.

Then Assumption 1 holds under either of the following sets of conditions.

- *(i)* ***Positive-semidefinite kernel.*** *Suppose that the centered kernel $h-\theta$ is positive semidefinite, its associated RKHS is separable, $\sup_x\{h(x,x)-\theta\} < \infty$, and for some $q > 1$, $\lambda_\ell \leq C\ell^{-q}$. If $0 < a < 1/2$, Assumption 1 holds for every $0 < \iota < \min\{1/2-a,\ a(q-1)\}$.*

(ii) ***Signed kernel.*** *Suppose that $\mathcal{X}$ is a separable metric space and $h - \theta$ is continuous at (x, x) for P-almost every x. Suppose that the spectral representation in* (2) *satisfies* $\sup_{\ell \geq 1} \|\psi_\ell\|_\infty < \infty$ *and for some* $\vartheta > 0$, $\sum_{\ell > R} |\lambda_\ell| = \mathcal{O}\left(R^{-\vartheta}\right)$ *as* $R \to \infty$. *If* $0 < a < \vartheta/(2\vartheta + 1)$, *Assumption 1 holds for every* $0 < \iota < \min\left\{\vartheta/(2\vartheta + 1) - a,\, a\vartheta\right\}$.

We first establish the empirical spectral approximation used in the proof of Proposition C.2. Throughout this section, let $K(x, y) := h(x, y) - \theta$, and define the oracle-centered normalized Gram matrix

$$\widetilde{\mathbf{K}}_n := \frac{1}{n}\{K(X_i, X_j)\}_{i,j=1}^n.$$

Its empirical counterpart used in Proposition C.2 is

$$\widehat{\mathbf{K}}_n := \frac{1}{n}\{h(X_i, X_j) - U_n\}_{i,j=1}^n.$$

For two spectra $\boldsymbol{a} = (a_\ell)_{\ell \geq 1}$ and $\boldsymbol{b} = (b_\ell)_{\ell \geq 1}$, padded by zeros when necessary, define $\delta_\infty(\boldsymbol{a}, \boldsymbol{b}) := \inf_\pi \sup_{\ell \geq 1} |a_\ell - b_{\pi(\ell)}|$, where the infimum is taken over spectral matchings. For a signed spectrum, positive and negative eigenvalues are matched separately.

Lemma C.1 (Empirical spectral approximation)**.** *Let $\mathcal{K}$ be the integral operator induced by K, with spectrum $(\lambda_\ell)_{\ell \geq 1}$.*

(i) Suppose that K is positive semidefinite with a separable RKHS and $\sup_x K(x, x) < \infty$. Then

$$\delta_\infty\left\{\lambda(\widetilde{\mathbf{K}}_n), \lambda(\mathcal{K})\right\} = \mathcal{O}_{a.s.}\left(\sqrt{\frac{\log n}{n}}\right).$$

(ii) Suppose that $\mathcal{X}$ is a separable metric space and K is continuous at (x, x) for P-almost every x. Suppose that $K(x, y) = \sum_{r \geq 1} \lambda_r \psi_r(x)\psi_r(y)$, $\sup_{r \geq 1} \|\psi_r\|_\infty \leq B$, and, for some $\vartheta > 0$, $\sum_{r > R} |\lambda_r| = \mathcal{O}(R^{-\vartheta})$, as $R \to \infty$. Then, for any deterministic sequence $R_n \to \infty$ satisfying $R_n \log(nR_n)/n \to 0$, we have

$$\delta_\infty\left\{\lambda(\widetilde{\mathbf{K}}_n), \lambda(\mathcal{K})\right\} = \mathcal{O}_{a.s.}\left\{\sqrt{\frac{R_n \log(nR_n)}{n}} + R_n^{-\vartheta}\right\}.$$

In particular, taking $R_n \asymp (n/\log n)^{1/(2\vartheta+1)}$ gives

$$\delta_\infty\left\{\lambda(\widetilde{\mathbf{K}}_n), \lambda(\mathcal{K})\right\} = \mathcal{O}_{a.s.}\left\{\left(\frac{\log n}{n}\right)^{\vartheta/(2\vartheta+1)}\right\}.$$

Proof. We prove the two cases separately.

Case (i): positive-semidefinite kernel. Let $\mathcal{H}$ denote the RKHS associated with K, and write $\Phi(x) = K(x, \cdot) \in \mathcal{H}$. Measurability of K and separability of $\mathcal{H}$ make Φ strongly measurable; the diagonal bound gives $\|\Phi(X)\|_{\mathcal{H}}^2 = K(X, X) < \infty$ with a uniform bound. Consider the population and empirical covariance operators

$$\mathcal{C} = \mathbb{E}\{\Phi(X) \otimes \Phi(X)\}, \qquad \mathcal{C}_n = \frac{1}{n}\sum_{i=1}^n \Phi(X_i) \otimes \Phi(X_i).$$

The nonzero eigenvalues of $\mathcal{C}$ coincide with those of $\mathcal{K}$, while the nonzero eigenvalues of $\mathcal{C}_n$ coincide with those of $\widetilde{\mathbf{K}}_n$. Here $\|\cdot\|_{\mathrm{HS}}$ denotes the Hilbert–Schmidt norm on operators on $\mathcal{H}$.

Since

$$\|\Phi(X) \otimes \Phi(X)\|_{\mathrm{HS}} = \|\Phi(X)\|_{\mathcal{H}}^2 = K(X, X) \leq \sup_x K(x, x) < \infty,$$

a Hilbert-space concentration inequality followed by the Borel–Cantelli lemma yields

$$\|\mathcal{C}_n - \mathcal{C}\|_{\mathrm{HS}} = \mathcal{O}_{a.s.}\left(\sqrt{\frac{\log n}{n}}\right).$$

By the Hoffman–Wielandt inequality for compact self-adjoint operators,

$$\delta_\infty \left\{ \lambda(\widetilde{\mathbf{K}}_n), \lambda(\mathcal{K}) \right\} \le \|\mathcal{C}_n - \mathcal{C}\|_{\mathrm{HS}},$$

which proves part (i).

Case (ii): signed kernel. For an integer $R \ge 1$, define the rank-R truncation

$$K_R(x,y) = \sum_{r=1}^{R} \lambda_r \psi_r(x)\psi_r(y), \qquad K_R^{\perp}(x,y) = K(x,y) - K_R(x,y).$$

The spectral-tail condition implies $\sum_r |\lambda_r| < \infty$. By Theorems 2.4 and 3.1 of [4], the diagonal continuity assumption therefore gives $K(X,X) = \sum_{r\ge1} \lambda_r \psi_r(X)^2$ almost surely. Together with the off-diagonal spectral expansion and the eigenfunction bound, this yields, simultaneously for all sample pairs, including $i = j$,

$$|K_R^{\perp}(X_i, X_j)| \le B^2 \sum_{r>R} |\lambda_r| = \mathcal{O}(R^{-\vartheta}) \quad \text{a.s.}$$

Hence

$$\left\| \frac{1}{n} \{K_R^{\perp}(X_i, X_j)\}_{i,j=1}^{n} \right\|_{\mathrm{op}} = \mathcal{O}(R^{-\vartheta}),$$

and, similarly,

$$\|\mathcal{K} - \mathcal{K}_R\|_{\mathrm{op}} \le \sup_{r>R} |\lambda_r| \le \sum_{r>R} |\lambda_r| = \mathcal{O}(R^{-\vartheta}).$$

It remains to control the empirical error of the finite-rank component. Let $\Psi_R = \{\psi_r(X_i)\}_{1\le i\le n,\ 1\le r\le R}$, and $G_{n,R} = \Psi_R^{\top}\Psi_R/n$. Since $(\psi_r)_{r\ge1}$ is orthonormal in $L^2(P)$, $\mathbb{E}G_{n,R} = I_R$. Moreover, $\sum_{r=1}^{R} \psi_r(X_i)^2 \le B^2 R$. A matrix Bernstein inequality and the Borel–Cantelli lemma therefore give

$$\|G_{n,R} - I_R\|_{\mathrm{op}} = \mathcal{O}_{a.s.}\left\{ \sqrt{\frac{R\log(nR)}{n}} + \frac{R\log(nR)}{n} \right\}.$$

Whenever $R\log(nR)/n \to 0$, the second term is of smaller order, so

$$\|G_{n,R} - I_R\|_{\mathrm{op}} = \mathcal{O}_{a.s.}\left\{ \sqrt{\frac{R\log(nR)}{n}} \right\}.$$

Let $D_R = \mathrm{diag}(\lambda_1, \dots, \lambda_R)$. The nonzero eigenvalues of the normalized empirical Gram matrix associated with K_R coincide with those of $G_{n,R}^{1/2} D_R G_{n,R}^{1/2}$. Since $\|G_{n,R} - I_R\|_{\mathrm{op}} = o_{a.s.}(1)$, $\|G_{n,R}^{1/2} - I_R\|_{\mathrm{op}} \lesssim \|G_{n,R} - I_R\|_{\mathrm{op}}$ eventually almost surely. Since $\|D_R\|_{\mathrm{op}} \le \sum_{r\ge1} |\lambda_r| < \infty$,

$$\begin{aligned} \|G_{n,R}^{1/2} D_R G_{n,R}^{1/2} - D_R\|_{\mathrm{op}} &\le \|(G_{n,R}^{1/2} - I_R) D_R G_{n,R}^{1/2}\|_{\mathrm{op}} + \|D_R (G_{n,R}^{1/2} - I_R)\|_{\mathrm{op}} \\ &= \mathcal{O}_{a.s.}\left\{ \sqrt{R\log(nR)/n} \right\}. \end{aligned}$$

Weyl's inequality therefore gives the same rate for the spectral matching error of the rank-R empirical and population operators. Combining this bound with the empirical and population truncation errors yields

$$\delta_\infty \left\{ \lambda(\widetilde{\mathbf{K}}_n), \lambda(\mathcal{K}) \right\} = \mathcal{O}_{a.s.}\left\{ \sqrt{\frac{R\log(nR)}{n}} + R^{-\vartheta} \right\}.$$

Finally, choosing $R_n \asymp (n/\log n)^{1/(2\vartheta+1)}$ balances the two terms and gives

$$\delta_\infty \left\{ \lambda(\widetilde{\mathbf{K}}_n), \lambda(\mathcal{K}) \right\} = \mathcal{O}_{a.s.}\left\{ \left(\frac{\log n}{n} \right)^{\vartheta/(2\vartheta+1)} \right\}.$$

□

Proof of Proposition C.2. We first transfer the oracle-centered spectral bounds in Lemma C.1 to the empirical centering used in the proposition. Since

$$\widehat{\mathbf{K}}_n = \widetilde{\mathbf{K}}_n - (U_n - \theta)\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}, \quad \left\|\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}\right\|_{\mathrm{op}} = 1,$$

Weyl's inequality gives $\delta_\infty\left\{\lambda(\widehat{\mathbf{K}}_n), \lambda(\widetilde{\mathbf{K}}_n)\right\} \le |U_n - \theta|$. Under the conditions of Theorem 4,

$$|U_n - \theta| = \mathcal{O}_{a.s.}\left(\frac{\log\log n}{n}\right),$$

which is negligible relative to either rate in Lemma C.1. Consequently, after the spectral matching described above,

$$\sup_{\ell\ge 1}|\widehat{\lambda}_{\ell,n} - \lambda_\ell| = \begin{cases} \mathcal{O}_{a.s.}\left(\sqrt{\log n/n}\right), & \text{under (i)}, \\ \mathcal{O}_{a.s.}\left\{(\log n/n)^{\vartheta/(2\vartheta+1)}\right\}, & \text{under (ii)}. \end{cases} \tag{17}$$

We next control the three positive spectral functionals. Since $\beta_\ell \asymp \ell^{-b}$ with $b > 1$,

$$\log(1/\beta_\ell) \lesssim \log(\ell+1).$$

Moreover, as shown in the proof of Theorem 5,

$$\{g^{-1}(\alpha\beta_\ell)\}^2 \lesssim 1 + \log(1/\beta_\ell) \lesssim \log(\ell+1).$$

Since $x \mapsto x_+$ is 1-Lipschitz, the three estimation errors are therefore bounded, up to a constant, by

$$L_n\log(L_n+1)\sup_{\ell\le L_n}|\widehat{\lambda}_{\ell,n} - \lambda_\ell| + \sum_{\ell > L_n}(\lambda_\ell)_+\log(\ell+1). \tag{18}$$

For case (i), the assumption $\lambda_\ell \le C\ell^{-q}$ with $q > 1$ gives

$$\sum_{\ell > L}\lambda_\ell\log(\ell+1) = \mathcal{O}\left\{L^{-(q-1)}\log L\right\}.$$

Since $L_n \asymp n^a$, (17)–(18) imply

$$\begin{aligned}&|\widehat{\Lambda}^+ - \Lambda^+| + |\widehat{\Lambda}^+_\beta - \Lambda^+_\beta| + |\widehat{\Lambda}^+_{\beta,g} - \Lambda^+_{\beta,g}| \\ &\quad = \mathcal{O}_{a.s.}\left\{n^{a-1/2}(\log n)^{3/2} + n^{-a(q-1)}\log n\right\}.\end{aligned}$$

Hence, for every

$$0 < \iota < \min\left\{\frac{1}{2} - a,\ a(q-1)\right\},$$

the preceding display is $\mathcal{O}_{a.s.}(n^{-\iota})$.

For case (ii), the spectral-tail assumption gives

$$\sum_{\ell > L}|\lambda_\ell|\log(\ell+1) = \mathcal{O}\left(L^{-\vartheta}\log L\right). \tag{19}$$

Indeed, by a dyadic decomposition,

$$\begin{aligned}\sum_{\ell > L}|\lambda_\ell|\log(\ell+1) &\le \sum_{j\ge 0}\log(2^{j+1}L+1)\sum_{\ell > 2^jL}|\lambda_\ell| \\ &\lesssim L^{-\vartheta}\sum_{j\ge 0}2^{-j\vartheta}\{\log L + j + 1\} \\ &= \mathcal{O}(L^{-\vartheta}\log L).\end{aligned}$$

Combining (17)–(19) with $L_n \asymp n^a$ gives

$$\begin{aligned}&|\widehat{\Lambda}^+ - \Lambda^+| + |\widehat{\Lambda}^+_\beta - \Lambda^+_\beta| + |\widehat{\Lambda}^+_{\beta,g} - \Lambda^+_{\beta,g}| \\ &= \mathcal{O}_{a.s.}\left(n^{a-\vartheta/(2\vartheta+1)}(\log n)^{1+\vartheta/(2\vartheta+1)} + n^{-a\vartheta}\log n\right).\end{aligned}$$

Therefore, for every

$$0 < \iota < \min\left\{\frac{\vartheta}{2\vartheta+1} - a,\ a\vartheta\right\},$$

these three errors are $\mathcal{O}_{a.s.}(n^{-\iota})$.

It remains to control $\widehat{\Lambda}$. In case (i), the separable RKHS construction and the diagonal bound give $\mathrm{tr}(\mathcal{C}) = \mathbb{E}\|\Phi(X)\|_{\mathcal{H}}^2 = \mathbb{E}K(X,X)$; the nonzero spectra of $\mathcal{C}$ and $\mathcal{K}$ coincide. In case (ii), the spectral-tail condition makes $\mathcal{K}$ trace class, and Corollary 3.2 of [4] applies by diagonal continuity. Thus, in either case,

$$\Lambda = \sum_{\ell \geq 1} \lambda_\ell = \mathbb{E}\{h(X,X) - \theta\},$$

and hence

$$\widehat{\Lambda} - \Lambda = \frac{1}{n}\sum_{i=1}^{n}[K(X_i,X_i) - \mathbb{E}K(X,X)] - (U_n - \theta).$$

Under case (i), $K(X,X)$ is uniformly bounded by assumption. Under case (ii),

$$|K(X,X)| \leq B^2 \sum_{\ell \geq 1} |\lambda_\ell| < \infty \quad \text{a.s.}$$

Thus, in either case, a scalar concentration inequality together with Borel–Cantelli gives

$$\frac{1}{n}\sum_{i=1}^{n}[K(X_i,X_i) - \mathbb{E}K(X,X)] = \mathcal{O}_{a.s.}\left(\sqrt{\frac{\log n}{n}}\right).$$

Combining this with $|U_n - \theta| = \mathcal{O}_{a.s.}(\log\log n/n)$ yields

$$|\widehat{\Lambda} - \Lambda| = \mathcal{O}_{a.s.}\left(\sqrt{\frac{\log n}{n}}\right).$$

This term is $\mathcal{O}_{a.s.}(n^{-\iota})$ for every admissible ι in either case. Combining the preceding bounds proves Assumption 1. □

C.3 Asymptotic validity of the data-driven weights in SAGE

Proposition C.3. *Let $\widehat{\lambda}_{\ell,n}$ and $\widehat{\Lambda}$ be the empirical spectral quantities defined in Proposition C.2. Suppose that $\Lambda^+ > 0$. For some $0 < \iota < \kappa \leq 1/2$, assume that*

$$|\widehat{\Lambda} - \Lambda| + \sup_{\ell \geq 1}|\widehat{\lambda}_{\ell,n} - \lambda_\ell| = \mathcal{O}_{a.s.}\left\{\left(\frac{\log n}{n}\right)^{\kappa}\right\}, \quad L_n(\log n)^{\kappa+1} = o(n^{\kappa-\iota}),$$

and

$$\sum_{\substack{\ell > L_n \\ \lambda_\ell > 0}} \lambda_\ell \left\{1 + \log\left(\frac{\Lambda^+}{\lambda_\ell}\right)\right\} = o(n^{-\iota}).$$

Then,

$$\sum_{\ell \leq L_n} |\widehat{\beta}_{\ell,n} - \beta_\ell^\star| + \sum_{\ell > L_n} \beta_\ell^\star = o_{a.s.}(n^{-\iota}).$$

Moreover, the spectral quantities computed using $\widehat{\beta}_{\ell,n}$ satisfy

$$|\widehat{\Lambda} - \Lambda| + |\widehat{\Lambda}^+ - \Lambda^+| + |\widehat{\Lambda}^+_{\widehat{\beta}} - \Lambda^+_{\beta^\star}| + |\widehat{\Lambda}^+_{\widehat{\beta},g} - \Lambda^+_{\beta^\star,g}| = o_{a.s.}(n^{-\iota}),$$

where $\widehat{\Lambda}^+_{\widehat{\beta}}$ and $\widehat{\Lambda}^+_{\widehat{\beta},g}$ are defined analogously to $\widehat{\Lambda}^+_{\beta}$ and $\widehat{\Lambda}^+_{\beta,g}$ in Proposition C.2, with β replaced by $\widehat{\beta}_n$.

Proof. Let

$$\delta_n := \sup_{\ell \geq 1} |\widehat{\lambda}_{\ell,n} - \lambda_\ell| = \mathcal{O}_{a.s.}\left\{ \left(\frac{\log n}{n} \right)^{\kappa} \right\}.$$

Since $x \mapsto x_+$ is 1-Lipschitz,

$$|\widehat{\Lambda}^+ - \Lambda^+| \leq \sum_{\ell \leq L_n} |(\widehat{\lambda}_{\ell,n})_+ - (\lambda_\ell)_+| + \sum_{\ell > L_n} (\lambda_\ell)_+ \leq L_n \delta_n + \sum_{\ell > L_n} (\lambda_\ell)_+ = o_{a.s.}(n^{-\iota}),$$

where the last equality follows from the two conditions on L_n and the positive spectral tail. Since $\Lambda^+ > 0$, it follows that $\widehat{\Lambda}^+$ is eventually bounded away from zero almost surely.

We first compare the empirical allocation with the fixed oracle allocation $\beta^\star_\ell = (\lambda_\ell)_+/\Lambda^+$. For $\ell \leq L_n$, $\widehat{\beta}_{\ell,n} = (\widehat{\lambda}_{\ell,n})_+/\widehat{\Lambda}^+$. Hence, using the preceding bound and $\sum_{\ell \leq L_n} (\widehat{\lambda}_{\ell,n})_+ = \widehat{\Lambda}^+$,

$$\begin{aligned} \sum_{\ell \leq L_n} |\widehat{\beta}_{\ell,n} - \beta^\star_\ell| &\leq \frac{1}{\Lambda^+} \sum_{\ell \leq L_n} |(\widehat{\lambda}_{\ell,n})_+ - (\lambda_\ell)_+| + \frac{|\widehat{\Lambda}^+ - \Lambda^+|}{\Lambda^+} \\ &\lesssim L_n \delta_n + \sum_{\ell > L_n} (\lambda_\ell)_+ = o_{a.s.}(n^{-\iota}). \end{aligned}$$

Moreover,

$$\sum_{\ell > L_n} \beta^\star_\ell = \frac{1}{\Lambda^+} \sum_{\ell > L_n} (\lambda_\ell)_+ = o(n^{-\iota}),$$

which proves the first assertion.

It remains to control the two weighted spectral quantities. We treat the entropy and Gaussian-mixture terms simultaneously. For the entropy term, set

$$F_s(x) := x \log(s/x), \qquad F_s(0) := 0.$$

On any compact interval of s bounded away from zero,

$$|F_s(x) - F_s(y)| \lesssim |x - y| \log\left(\frac{e}{|x - y|} \right),$$

uniformly for x, y in a bounded interval. Since $\widehat{\Lambda}^+ \to \Lambda^+ > 0$ almost surely, splitting the empirical–oracle difference into the perturbation of the first L_n eigenvalues, the normalization error $\widehat{\Lambda}^+ - \Lambda^+$, and the truncated population tail gives

$$|\widehat{\Lambda}^+_{\widehat{\beta}} - \Lambda^+_{\beta^\star}| \lesssim_{a.s.} \; L_n \delta_n \log(e/\delta_n) + |\widehat{\Lambda}^+ - \Lambda^+| + \sum_{\ell > L_n} (\lambda_\ell)_+ \log\left(\frac{\Lambda^+}{(\lambda_\ell)_+} \right).$$

Because $\log(e/\delta_n) = \mathcal{O}(\log n)$ almost surely, the first term is

$$\mathcal{O}_{a.s.}\left\{ L_n n^{-\kappa} (\log n)^{\kappa+1} \right\} = o_{a.s.}(n^{-\iota}),$$

while the remaining terms are of the same order by the assumptions. Thus

$$|\widehat{\Lambda}^+_{\widehat{\beta}} - \Lambda^+_{\beta^\star}| = o_{a.s.}(n^{-\iota}).$$

For the Gaussian-mixture term, define

$$F_s(x) := x \left\{ g^{-1}\left(\frac{\alpha x}{s} \right) \right\}^2, \qquad F_s(0) := 0.$$

The tail behavior of g^{-1} established in the proof of Theorem 5 yields, uniformly for $0 < x \leq s$ with s in a neighborhood of Λ^+,

$$F_s(x) \lesssim x \left\{ 1 + \log(s/x) \right\}, \qquad |F_s(x) - F_s(y)| \lesssim |x - y| \log\left(\frac{e}{|x - y|} \right).$$

Moreover, the dependence on s is locally Lipschitz after summation. Indeed, writing $q(u) = \{g^{-1}(u)\}^2$, we have $q'(u) = -1/\{g^{-1}(u)\phi(g^{-1}(u))\}$ and hence $u|q'(u)| \lesssim 1$ for $0 < u \leq$

α. Consequently, the normalization perturbation contributes only $\mathcal{O}_{a.s.}(|\widehat{\Lambda}^+ - \Lambda^+|)$. The same decomposition as above therefore gives

$$|\widehat{\Lambda}^+_{\widehat{\beta},g} - \Lambda^+_{\beta^\star,g}| \lesssim_{a.s.} L_n \delta_n \log(e/\delta_n) + |\widehat{\Lambda}^+ - \Lambda^+| + \sum_{\ell > L_n} (\lambda_\ell)_+ \left\{ 1 + \log\left(\frac{\Lambda^+}{(\lambda_\ell)_+} \right) \right\},$$

and hence

$$|\widehat{\Lambda}^+_{\widehat{\beta},g} - \Lambda^+_{\beta^\star,g}| = o_{a.s.}(n^{-\iota}).$$

Finally, since $\iota < \kappa$,

$$|\widehat{\Lambda} - \Lambda| = \mathcal{O}_{a.s.} \left\{ \left(\frac{\log n}{n} \right)^\kappa \right\} = o_{a.s.}(n^{-\iota}).$$

Combining this with the preceding bounds yields

$$|\widehat{\Lambda} - \Lambda| + |\widehat{\Lambda}^+ - \Lambda^+| + |\widehat{\Lambda}^+_{\widehat{\beta}} - \Lambda^+_{\beta^\star}| + |\widehat{\Lambda}^+_{\widehat{\beta},g} - \Lambda^+_{\beta^\star,g}| = o_{a.s.}(n^{-\iota}),$$

so the data-driven spectral quantities satisfy Assumption 1 relative to the fixed oracle allocation $\beta^\star$. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Introduction state the two regimes, the SAGE boundary, plug-in spectral calibration, and the role of the simulations; Sections 3, 4, and 5 contain the corresponding technical results, examples, and empirical checks. The coverage and width claims are stated as asymptotic guarantees under explicit assumptions rather than as finite-sample guarantees.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Conclusion discusses the main limitation of the degenerate implementation, namely the need for a verified spectral approximation rate and the use of multiplicative finite-sample inflation. Section 5 also reports finite-sample sensitivity under heavy tails and delayed starts.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The manuscript contains theorem statements and proofs in the technical appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5 specifies the main simulation design, including α, m, $n_{\max}$, the number of Monte Carlo replications, data-generating distributions, statistics, and stopping rule.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The manuscript as provided includes an anonymized code in supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 describes the simulation horizon, delayed start, number of Monte Carlo replications, nominal level, compared intervals, data-generating distributions, target functionals, and optional-stopping rule. No training/test split or optimizer is involved because the experiments are simulation studies rather than model training experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experiments report empirical coverage and stopping behavior over $M = 500$ Monte Carlo replications.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experimental sections describe the number of replications and algorithmic settings. The simulation code is lightweight and uses standard scientific Python libraries.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is theoretical and simulation-based, uses no private or sensitive data, involves no human subjects, and does not release high-risk models or datasets. Based on the current manuscript, no deviation from the NeurIPS Code of Ethics is apparent.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: The paper is foundational statistical methodology for anytime-valid inference and is not tied to a deployed application or a specific user population. Its likely impact is indirect, through improved sequential uncertainty quantification, and the manuscript does not introduce an immediate societal deployment pathway.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release pretrained models, scraped datasets, human data, or other assets with high misuse or dual-use risk.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: The experiments use synthetic data generated from distributions described in Section 5. The paper does not use existing datasets, models, or code assets beyond cited prior literature.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The current manuscript does not introduce or release a new dataset, benchmark, model, or code package as an asset.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The work does not involve crowdsourcing, user studies, or human-subject experiments.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The work does not involve human-subject experiments or personally identifiable data, so IRB or equivalent review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as an important, original, or non-standard component of the core methodology or experiments. Writing, editing, or formatting assistance, if any, is outside the methodological contribution and does not require declaration under the stated checklist question.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.