Game-theoretic statistics & sequential anytime-valid inference (SAVI): a martingale theory of evidence

Aaditya Ramdas

Dept. of Statistics and Data Science
Machine Learning Dept.
Carnegie Mellon University

Acknowledge support: NSF CAREER, Google Research Award, Adobe Research Award

Outline of this tutorial

✓ First half: game-theoretic hypothesis testing

B. Second half: game-theoretic estimation

Slides and references at

http://www.stat.cmu.edu/~aramdas/icml25

Outline of second half

1. Core definition: confidence sequence
2. A simple, explicit nonparametric example
3. Asymptotic confidence sequences

A "confidence sequence (CS)" for a parameter θ is a sequence of confidence intervals (L_n, U_n) that are constructed from the first n samples, and have a uniform (simultaneous) coverage guarantee.

$$\mathbb{P}(\forall t \geq 1 : \theta \in (L_t, U_t)) \geq 1 - \alpha\,.$$

(Another motivation: (L_n, U_n) should not contradict (L_m, U_m) for any $m > n$.
+pointwise CIs, intersection $= \emptyset$ a.s.,
but +CSs, intersection $= \theta$ w.p. $1 - \alpha$)

Darling, Robbins '67, '70s
Lai '76, '84
Robbins, Siegmund '70s

Much stronger than the pointwise (fixed-sample) confidence interval (CI) guarantee:

$$\forall n \geq 1, \mathbb{P}(\theta \in (\tilde{L}_n, \tilde{U}_n)) \geq 1 - \alpha\,.$$

$$\mathbb{P}(\forall n \geq 1 : \theta \in (L_n, U_n)) \geq 1 - \alpha\,.$$

Equivalent definitions:

$$\mathbb{P}(\exists n \in \mathbb{N} : \theta \notin (L_n, U_n)) \leq \alpha\,.$$

$$\mathbb{P}\Big(\bigcup_{n \in \mathbb{N}} \{\theta \notin (L_n, U_n)\}\Big) \leq \alpha\,.$$

More generally:

$$\mathbb{P}(\forall n \geq n_0 : \theta_n \in C_n) \geq 1 - \alpha\,.$$

$$\mathbb{P}(\exists n \in 2^{\mathbb{N}} : \theta \notin (L_n, U_n)) \leq \alpha\,.$$

$$\mathbb{P}\Big(\bigcup_{n\in\mathbb{N}} \{\theta \notin (L_n, U_n)\}\Big) \leq \alpha\,.$$

Some implications:

1. Valid inference at arbitrary stopping times:

   For any stopping time τ : $\mathbb{P}(\theta \notin (L_\tau, U_\tau)) \leq \alpha\,.$

2. Valid post-hoc inference (in hindsight):

   For any random time T : $\mathbb{P}(\theta \notin (L_T, U_T)) \leq \alpha\,.$

3. No pre-specified sample size:
   can extend or stop experiments adaptively.

Fact: the aforementioned properties imply each other.

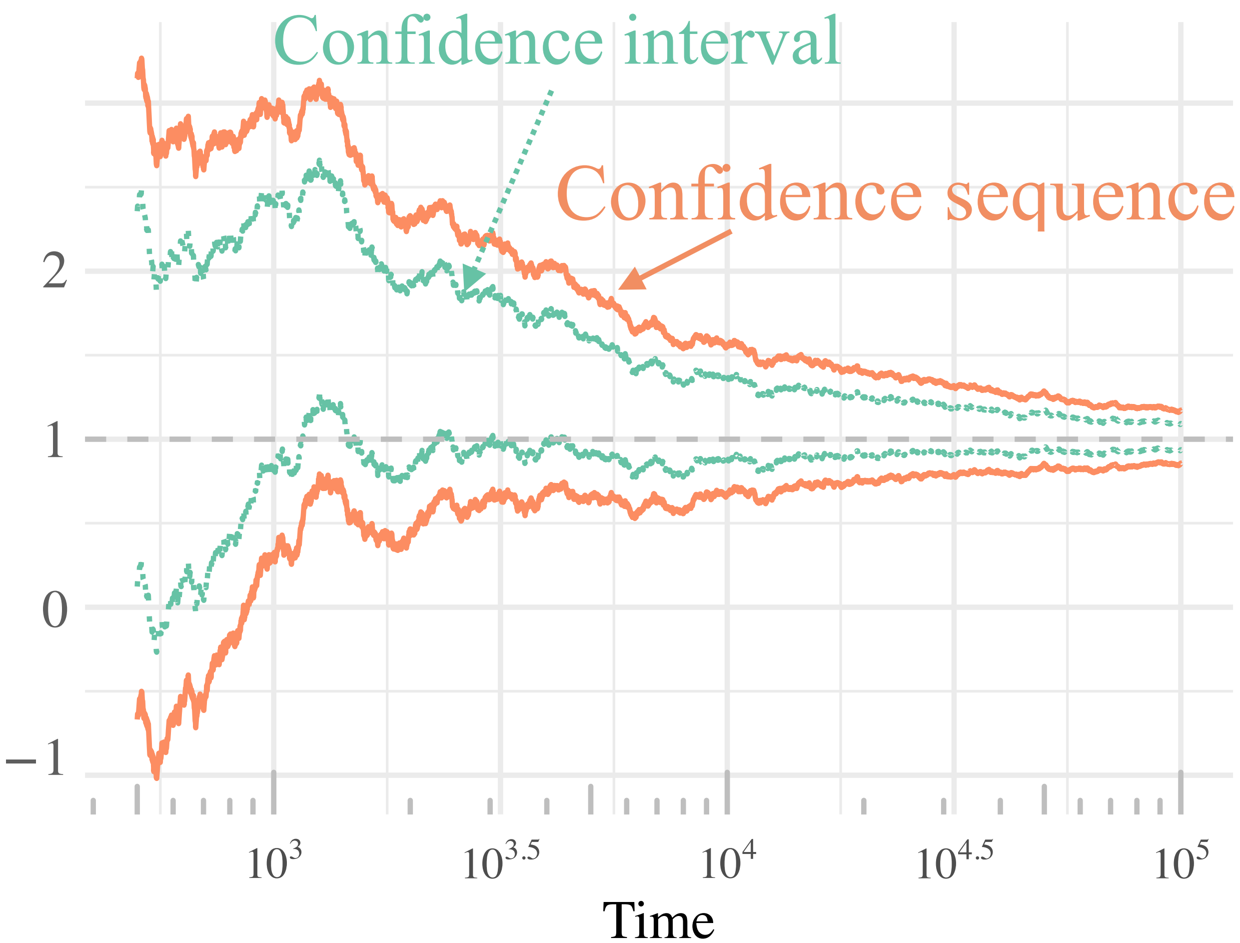
Eg: If $X_1, X_2, \ldots$ are iid Gaussian or bounded in $[-1,1]$,

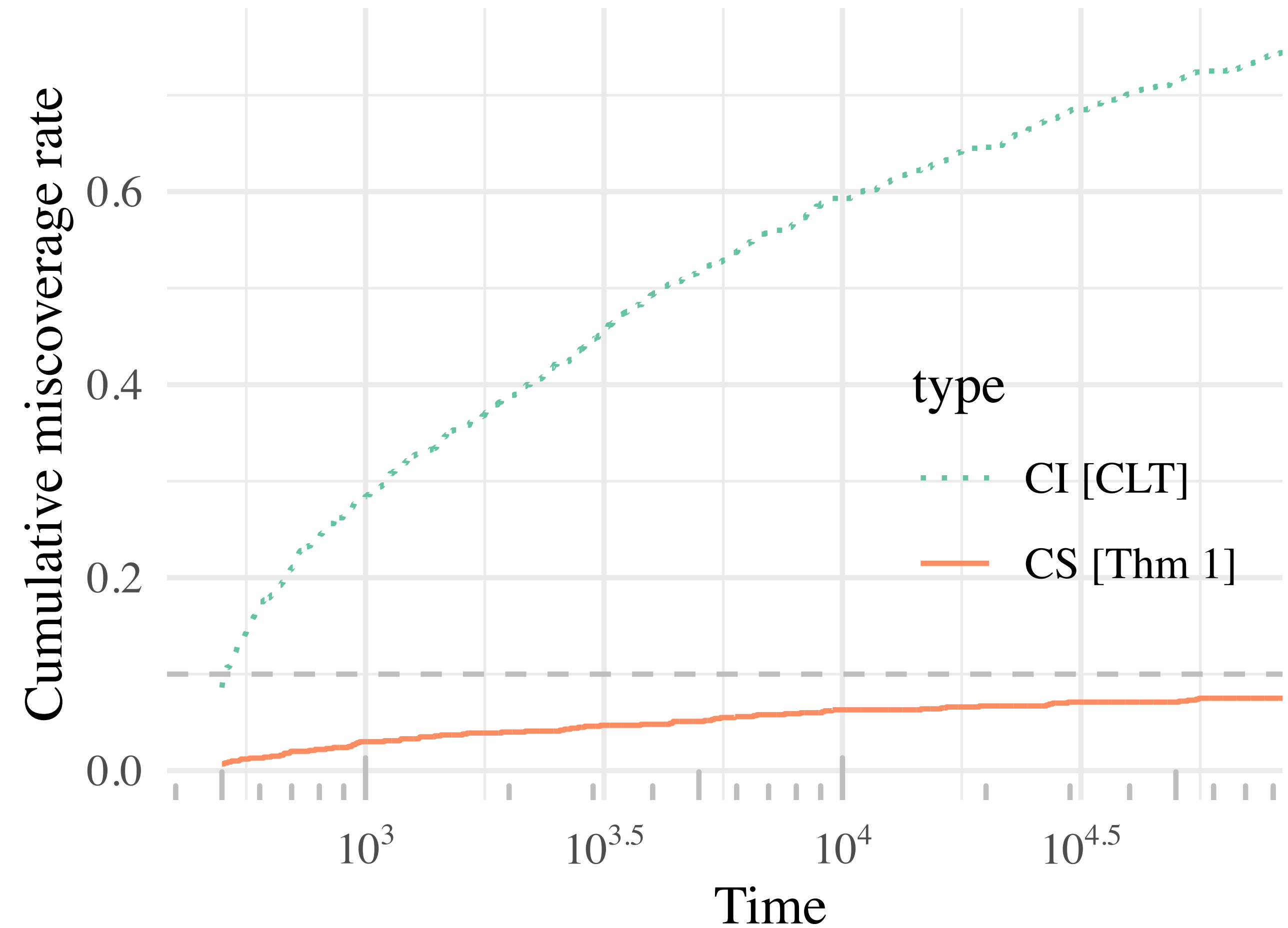
$$\frac{\sum_{i=1}^{n} X_i}{n} \pm 1.71\sqrt{\frac{\log\log(2n) + 0.72\log(10.4/\alpha)}{n}}$$

is a $(1-\alpha)$ confidence sequence for μ, as is

$$\bigcap_{s\le n} \frac{\sum_{i=1}^{s} X_i}{s} \pm 1.71\sqrt{\frac{\log\log(2s) + 0.72\log(10.4/\alpha)}{s}}.$$

Howard, Ramdas, McAuliffe, Sekhon '21





Confidence interval: $\forall n,\ \Pr(\theta \in \dot{C}_n) \geq (1-\alpha)\,.$

Confidence sequence: $\Pr(\forall n,\ \theta \in \bar{C}_n) \geq (1-\alpha)\,.$

$\iff\ \Pr(\theta \in \bar{C}_\tau) \geq (1-\alpha)$ for all stopping times τ

Confidence sequence for fixed quantiles

Define $u_t := \sqrt{\dfrac{0.73 \log\log(2.04t) + 0.52 \log(9.97/\alpha)}{t}}$

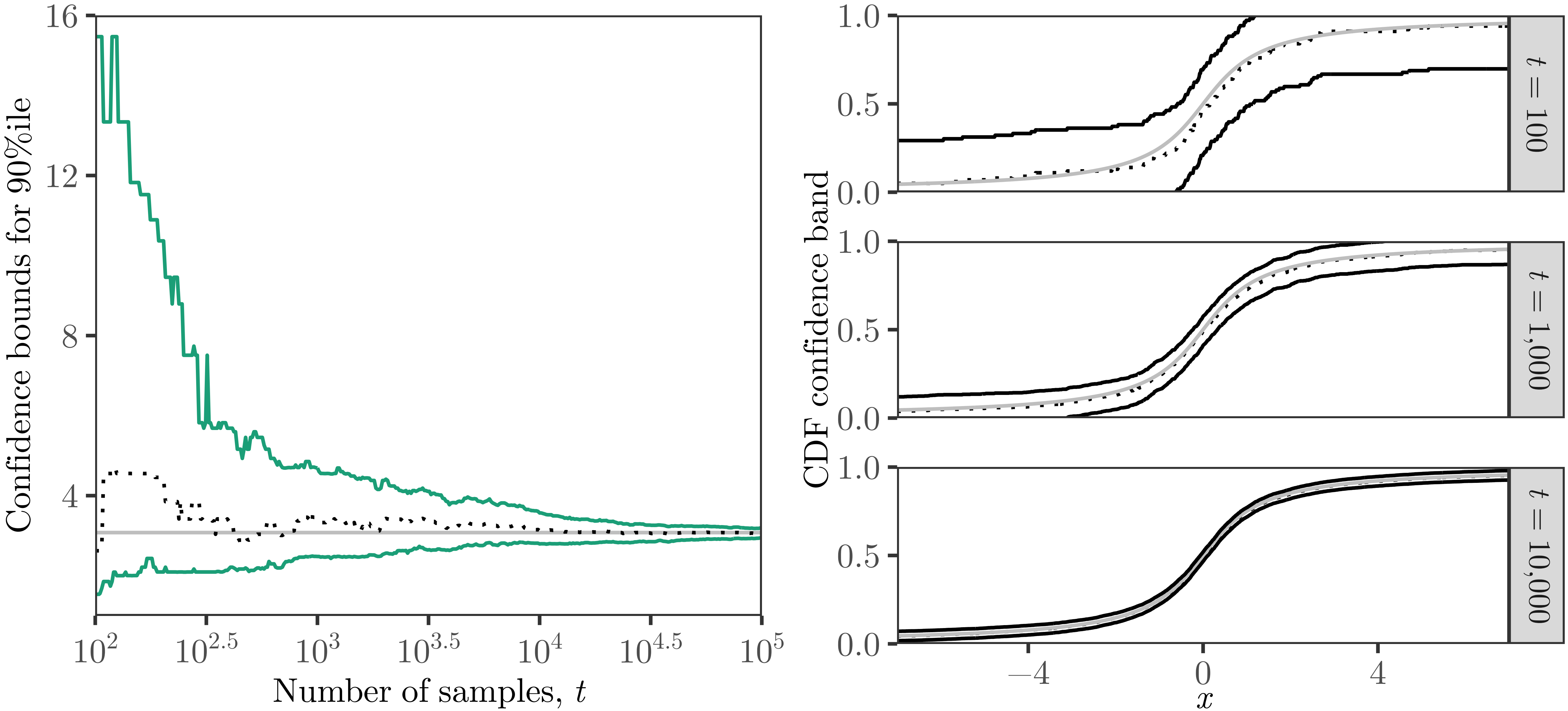
Then $\Pr(\forall t \in \mathbb{N} : \widehat{Q}_t(1/2 - u_t) \leq Q(1/2) \leq \widehat{Q}_t(1/2 + u_t)) \geq 1 - \alpha$.

Confidence sequence for all quantiles simultaneously

Define $u_t := \sqrt{\dfrac{\log\log(et) + 0.75 \log(34/\alpha)}{t}}$

$\Pr(\forall t \in \mathbb{N}, p \in (0,1) : \widehat{Q}_t(p - u_t) \leq Q(p) \leq \widehat{Q}_t(p + u_t)) \geq 1 - \alpha$.

Cauchy distribution



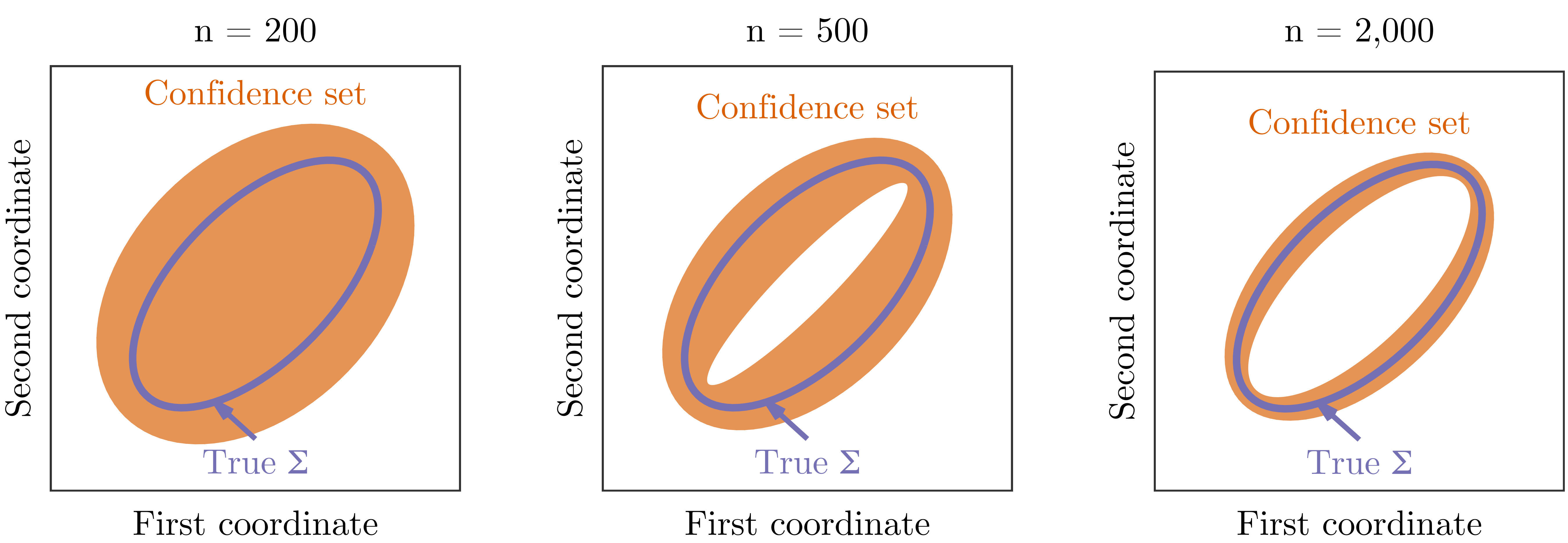
Only 0.9 quantile

All quantiles simultaneously

Eg: sequential covariance matrix estimation

Consider $X \in \mathbb{R}^d, EX = 0, |X_i| \leq b$.

$$\|\widehat{\Sigma}_n - \Sigma\|_{\mathrm{op}} \precsim \sqrt{\frac{b\log(d\log n)}{n}} + \frac{b\log(d\log n)}{n} \text{ uniformly w.h.p.}$$



Outline of second half

1. Core definition: confidence sequence
2. A simple, explicit nonparametric example
3. Asymptotic confidence sequences

Let $X_1, X_2, \ldots,$ be iid r.v. $\in [0,1]$, with mean μ.

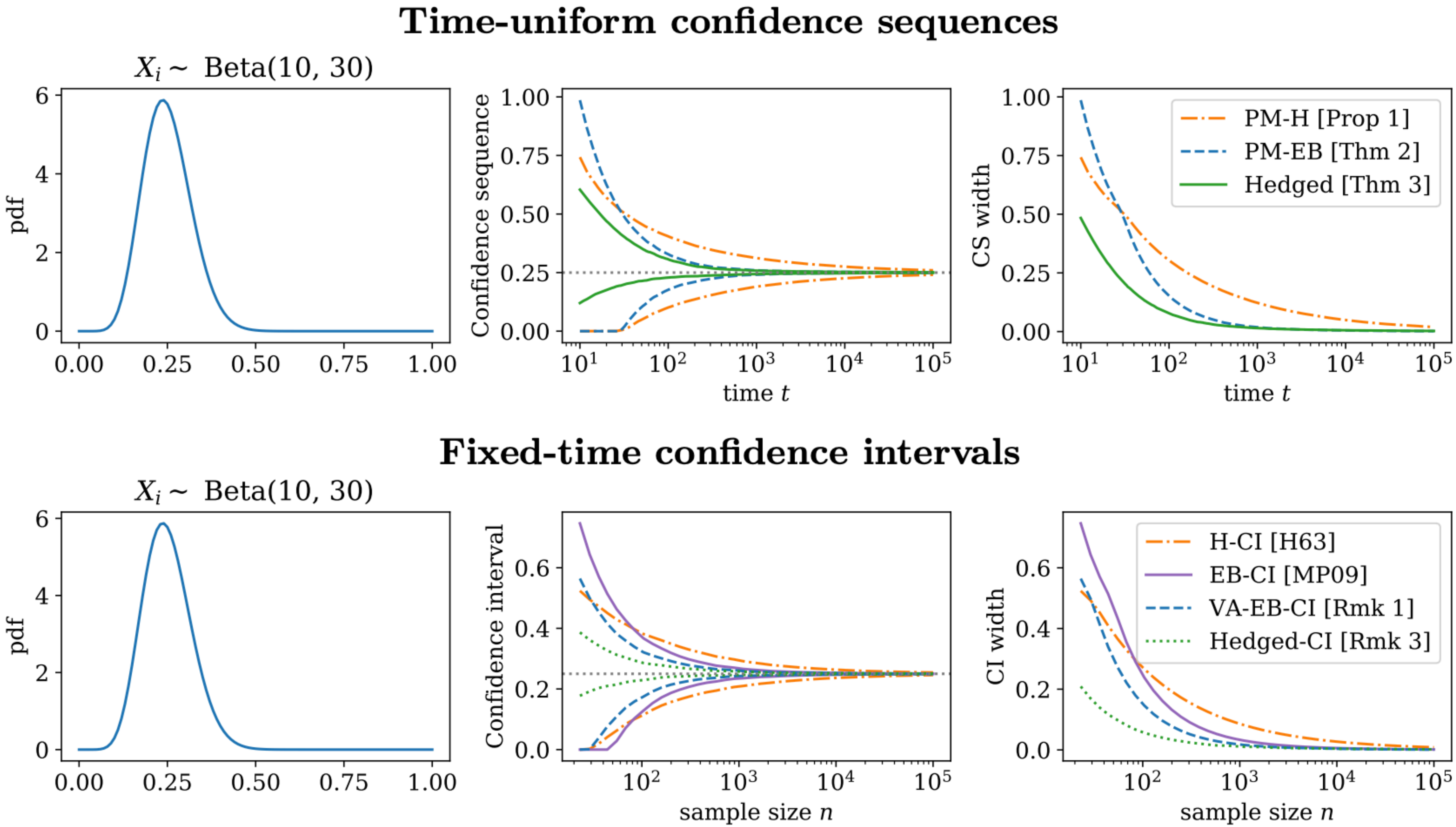
Q1. How can we construct a confidence interval for μ?

A1. Hoeffding: $\left[\overline{X}_n \pm \sqrt{\frac{\log(2/\alpha)}{2n}}\right] \cap [0,1]$

A2. Empirical Bernstein: $\left[\overline{X}_n \pm \sqrt{\frac{2\hat{\sigma}^2 \log(4/\alpha)}{n}} + \frac{7\log(4/\alpha)}{3(n-1)}\right]$

A3: Betting — significantly tighter!

Q2. How can we construct a confidence *sequence* for μ?



Time-uniform confidence sequences
$X_i \sim$ Beta(10, 30)
pdf
Confidence sequence
time t
CS width
PM-H [Prop 1]
PM-EB [Thm 2]
Hedged [Thm 3]
Fixed-time confidence intervals
$X_i \sim$ Beta(10, 30)
pdf
Confidence interval
sample size n
CI width
H-CI [H63]
EB-CI [MP09]
VA-EB-CI [Rmk 1]
Hedged-CI [Rmk 3]

Converting the problem to a game

Initial capital $K_0^{(m)} = 1$ for every (game) $m \in [0,1]$.

For each $t = 1,2,\ldots$

For each $m \in [0,1]$, statistician declares "bet" $\lambda_t^{(m)} \in \left[-\frac{1}{1-m}, \frac{1}{m}\right]$

Reality reveals X_t

Statistician's wealth in game m becomes $K_t^{(m)} = K_{t-1}^{(m)} \cdot (1 + \lambda_t^{(m)}(X_t - m))$

$$C_t := \left\{m \in [0,1] : K_t^{(m)} < 1/\alpha\right\}$$

(the games in which the statistician did not earn enough wealth)

Theorem: For any betting strategy, $(C_t)_{t\geq 1}$ is a confidence sequence for the true mean μ.

Two questions: Why is C_t a valid confidence sequence for μ? How do we bet so that it is an efficient (small) set?

1. For each $m \in [0,1]$, let us test $H_0^{(m)} : \mathbb{E}_P[X_i \mid X_1, \ldots, X_{i-1}] = m$.

$$K_t^{(m)} := \prod_{i \le t} (1 + \lambda_i^{(m)}(X_i - m)), \text{ where } \underbrace{\lambda_i^{(m)} \in [-1/(1-m), 1/m]}_{\text{predictable}}.$$

2. $C_t := \{m : K_t^{(m)} < 1/\alpha\}$ yields a confidence sequence for μ.

$$\sup_{P \in H_0^{(\mu)}} P(\exists t \in \mathbb{N} : \mu \notin C_t) \le \alpha .$$

1. For each $m \in [0,1]$, let us test $H_0^{(m)} : \mathbb{E}_P[X_i \mid X_1, \ldots, X_{i-1}] = m$.

$K_t^{(m)} := \prod_{i \leq t} (1 + \lambda_i^{(m)}(X_i - m))$, where $\lambda_i^{(m)} \in [-1/(1-m), 1/m]$.

$K_t^{(\mu)}$ is a nonnegative martingale +initial value one ("test martingale").

Ville's inequality $\sup_{P \in H_0^{(\mu)}} P(\exists t \in \mathbb{N} : K_t^{(\mu)} \geq 1/\alpha) \leq \alpha$.

C_t is incorrect only if $K_t^{(\mu)}$ exceeds $1/\alpha$. But this is happens w.p. $\leq \alpha$.

2. $C_t := \{m : K_t^{(m)} < 1/\alpha\}$ is a confidence sequence for μ.

$$\sup_{P \in H_0^{(\mu)}} P(\exists t \in \mathbb{N} : \mu \notin C_t) \leq \alpha .$$

Betting strategy 1: GRAPA

Growth Rate Adaptive to the Particular Alternative

$$\lambda_t^m(P) := \arg \max_{\lambda \in [-1/(1-m), 1/m]} \mathbb{E}_P[\log(1 + \lambda(X_t - m)) \mid \mathcal{F}_{t-1}] .$$

But we don't know P. <u>Approximate solution</u>:
differentiate wrt λ, and set equal to zero (KKT),
Taylor expand, and plug-in empirical estimates.

$$\lambda_t^m = \frac{\widehat{\mu}_t - m}{\widehat{\sigma}_t^2 + (\widehat{\mu}_t - m)^2} .$$

($\widehat{\mu}_t$ and $\widehat{\sigma}_t^2$ use the first $t-1$ samples)

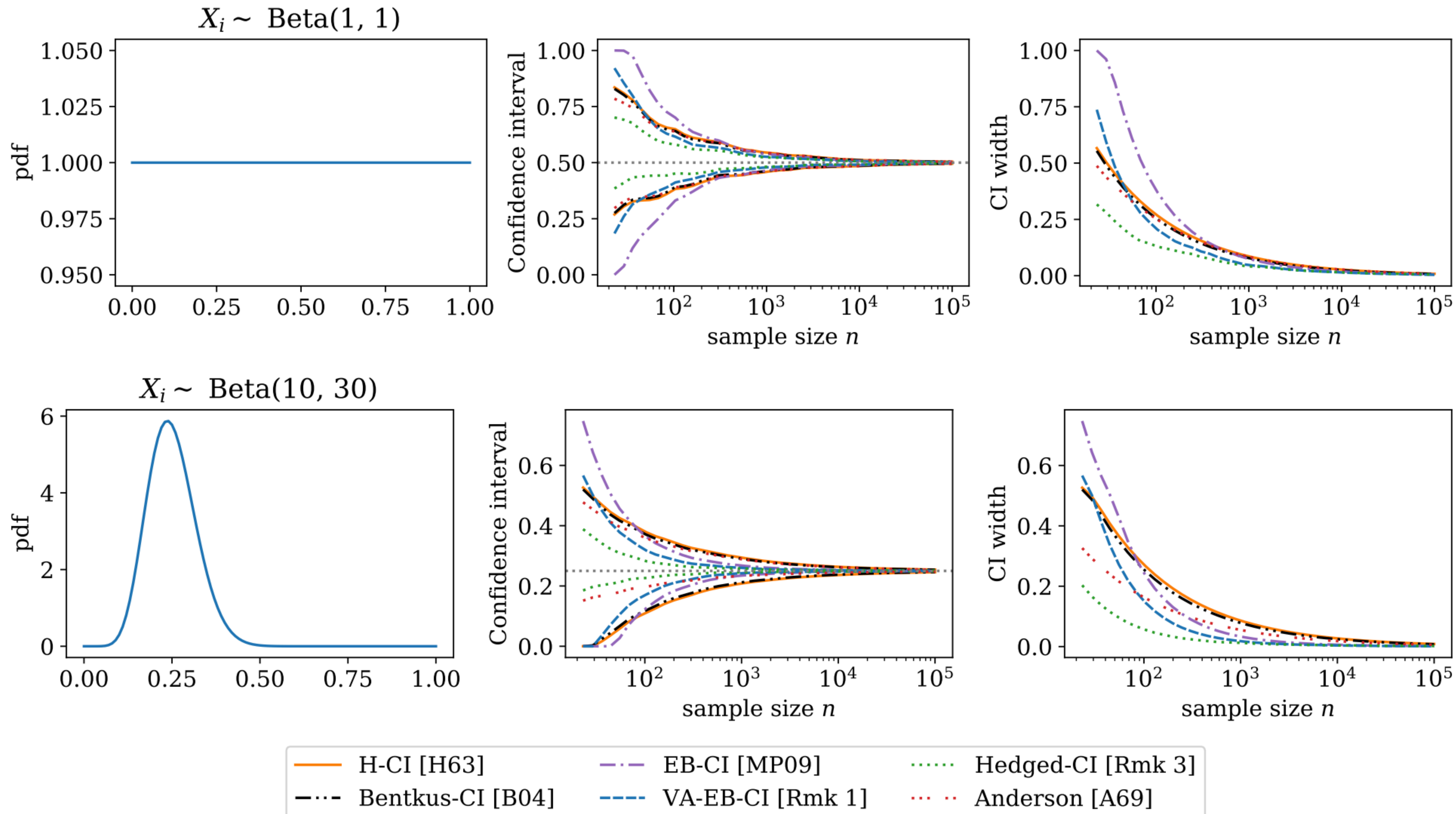
Betting strategy 2: Mixture

$$C_t := \left\{ m \in [0,1] : \int_{-1/(1-m)}^{1/m} \prod_{i=1}^{t} (1 + \lambda(X_i - m)) \, d\nu(\lambda) < 1/\alpha \right\}$$

In practice, we would discretize the mixture, does not affect validity.
We must make the mixture finer over time to preserve power.

We can use the connection to Cover's Portfolio Optimization
to analyze its performance (eg: uniform mixture will do).

Waudby-Smith+R (2024)
Jun+Orabona (2024)
Shekhar+R (2024)



In iid settings, $\lim_{n\to\infty} \sqrt{n}\text{Width}(C_n) - \sqrt{n}\text{Width}(\text{Bernstein}) \leq 0$,

(i.e. we match / beat the leading term of Bernstein's inequality, even though we do not know σ — tight "empirical Bernstein")

Shekhar + Ramdas (2023, arXiv)

The first sharp closed-form empirical Bernstein bound

Theorem 2 (Predictably-mixed empirical Bernstein CS **[PM-EB]**).
Suppose that $(X_t)_{t=1}^{\infty} \sim P$ *for some* $P \in \mathcal{P}^{\mu}$. *For any chosen* $(0,1)$*-valued predictable sequence* $(\lambda_t)_{t=1}^{\infty}$,

$$C_t^{\text{PM-EB}} := \left(\frac{\sum_{i=1}^t \lambda_i X_i}{\sum_{i=1}^t \lambda_i} \pm \frac{\log(2/\alpha) + \sum_{i=1}^t v_i \psi_E(\lambda_i)}{\sum_{i=1}^t \lambda_i} \right) \quad \textit{forms a } (1-\alpha)\textit{-CS for } \mu,$$

as does its running intersection, $\bigcap_{i \le t} C_i^{\text{PM-EB}}$.

$$\lim_{n\to\infty} \sqrt{n}\,\text{Width}(C_n) - \sqrt{n}\,\text{Width}(\text{Bernstein}) \to 0$$

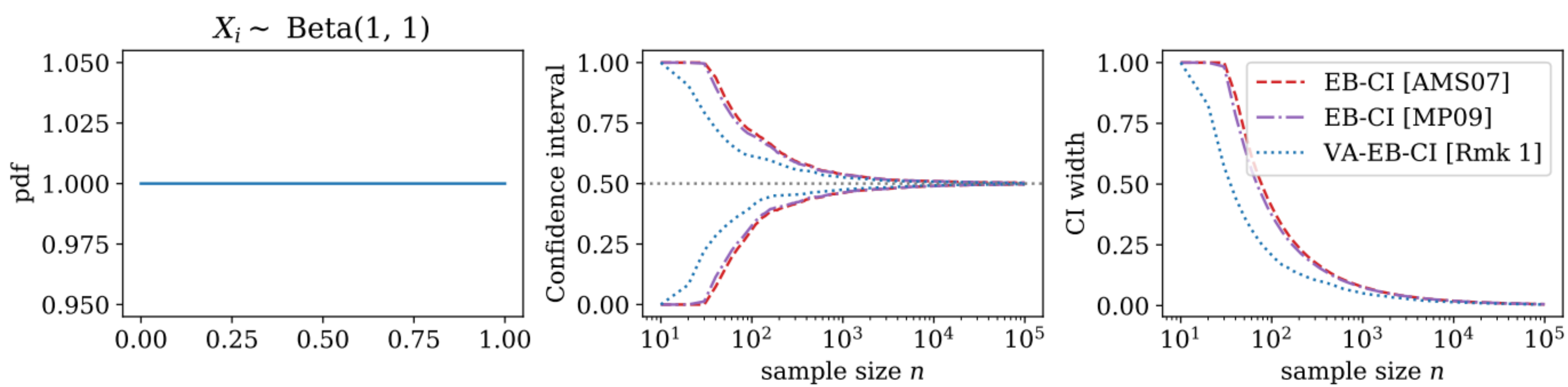


Figure 3: Comparison of the variance-adaptive empirical Bernstein CI with Maurer & Pontil's (MP09) and Audibert et al.'s (AMS07) empirical Bernstein CIs.

Aside: Kernel two-sample testing by betting (sequential MMD)

Observe $X_1, X_2, \ldots \sim P$ and $Y_1, Y_2, \ldots \sim Q$
$H_0 : P = Q$ versus $H_1 : P \neq Q$

Choose a kernel k, say bounded by one for simplicity.

Define $K_t = \prod_{i=1}^{t} (1 + \lambda_i [f_i(X_i) - f_i(Y_i)])$,

where f_i is a predictable function in the RKHS
(based on $X_1 \ldots X_{i-1}, Y_1, \ldots, Y_{i-1}$),
and λ_i is a predictable scalar in $[-1,1]$.

Eg: set $f_i \propto \sum_{j<i} \phi(X_j) - \phi(Y_j)$, or Online Gradient Descent in RKHS,
and pick λ using Online Newton Step, or universal portfolios.

$(K_t)_{t \geq 0}$ is a test martingale for H_0, and its growth rate under any alternative can be shown to be $\propto$ MMD(P, Q).

Shekhar+R (2023), Podkopaev et al. (2024)

Outline of second half

1. Core definition: confidence sequence
2. A simple, explicit nonparametric example
3. Asymptotic confidence sequences

Statistical problem	**Confidence interval**	**Confidence sequence**
Parametric inference	Wald, Neyman, Fisher	Robbins + co. (1967-76) Wasserman et al. (2020) Waudby-Smith & Ramdas (2020)
Sub-Gaussian mean estimation	Hoeffding (1963)	Robbins (1970) Howard et al. (2021)
Bounded mean estimation	Hoeffding (1963) Waudby-Smith & Ramdas (2024)	Howard et al. (2021) Waudby-Smith & Ramdas (2024)
Quantiles & CDFs	DKW (1956)	Howard & Ramdas (2021)
Sampling without replacement	Hoeffding (1963), Bardenet & Maillard (2015)	Waudby-Smith & Ramdas (2020, 2024)
Heavy-tailed mean estimation	Catoni (2012) Lugosi-Mendelson (2014+)	Wang & Ramdas (2023) Martinez-Taboada et al. (2025)
Nonasymptotic inference is impossible, but asymptotic inference is possible	Central limit theorem	?

Definition (AsympCS)

$(\widehat{\mu}_t \pm \bar{B}_t)_{t=1}^{\infty}$ is a $(1-\alpha)$-AsympCS for μ if there exists a *nonasymptotic* $(1-\alpha)$-CS for μ given by $(\widehat{\mu}_t \pm \bar{B}_t^\star)_{t=1}^{\infty}$, and

$$\bar{B}_t^\star / \bar{B}_t \to 1 \quad \text{almost surely.}$$

In words, an AsympCS is an arbitrarily precise a.s. approximation to a nonasymptotic CS for large t.

Why is this a sensible definition? The canonical CLT-based asymptotic *confidence interval* looks like

$$(\widehat{\mu}_n \pm \dot{B}_n) \quad \text{where} \quad \dot{B}_n := \widehat{\sigma}_n \frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{n}}$$

Fact: When invoking the CLT, there exists a nonasymptotic $\dot{B}_n^\star$ such that $\dot{B}_n^\star / \dot{B}_n \xrightarrow{P} 1$.

In contrast, our definition of AsympCSs requires $\bar{B}_t^\star / \bar{B}_t \xrightarrow{\text{a.s.}} 1.$

Theorem 1 (AsympCS for the mean of iid random variables)

Suppose $(Y_t)_{t=1}^{\infty} \overset{\text{iid}}{\sim} P$ with mean μ and finite variance. Then for any $\rho > 0$,

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

forms a $(1-\alpha)$-AsympCS for μ.

Paper has Lindeberg-Levy (non-iid, martingale) AsympCS

Theorem 2 (Asymptotic time-uniform coverage guarantees)

Suppose we tune $\rho_m = \sqrt{(-\log \alpha + \log(-2 \log \alpha) + 1)/m \widehat{\sigma}_m^2 \log m}$

and let $(C_t(m))_{t=1}^{\infty}$ be the AsympCS $+\rho_m$ in place of ρ. Then

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

As you start (at time m) later and later, the *time-uniform* type-I error approaches α. (This could have been an alternate definition of AsympCSs.)

AsympCI:

$$\liminf_{m\to\infty} \mathbb{P}(\mu \in \dot{C}_m) = 1-\alpha$$

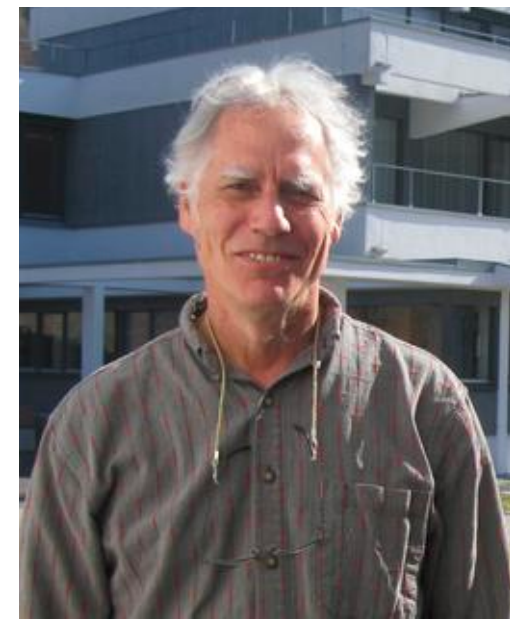
$$\iff \limsup_{m\to\infty} \mathbb{P}(\mu \notin \dot{C}_m) = \alpha$$

AsympCS:

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

$$\iff \limsup_{m\to\infty} \mathbb{P}(\exists t \geq m : \mu \notin \bar{C}_t(m)) = \alpha$$

Now that we have CSs under CLT-like assumptions, we can do doubly-robust causal inference in *sequential settings at stopping times.*



(or, Robbins meets Robins.)

Given $(X_t, A_t, Y_t)_{t=1}^{\infty} \sim P$, wish to estimate

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

- X — covariates (e.g. age, sex, etc.).
- A — treatment level (e.g. 1 for treatment, 0 for placebo).
- Y — outcome of interest (e.g. whether patient recovered from sickness).

Classical AIPW "doubly robust" estimator (Robins et al. 1994):

$$\hat{\psi}_t := \frac{1}{t} \sum_{i=1}^{t} \hat{f}_t(X_i, A_i, Y_i)$$

where $\hat{f}_t$ involves estimates $(\widehat{\mu}_t^1, \widehat{\mu}_t^0, \widehat{\pi}_t)$ of

regression functions

$$\mu^a(x) = \mathbb{E}(Y \mid X = x, A = a),$$

and the propensity score

$$\pi(x) = \Pr(A = 1 \mid X = x).$$

Theorem (AsympCS for the ATE)

Given observations $(X_t, A_t, Y_t)_{t=1}^{\infty} \sim P$, construct a (sequentially) cross-fit DR estimator $\widehat{\psi}_t^{\times}$. Suppose $\|\widehat{\mu}_t^a - \mu^a\| \|\widehat{\pi} - \pi\| = o(\sqrt{\log t / t})$.

Then,

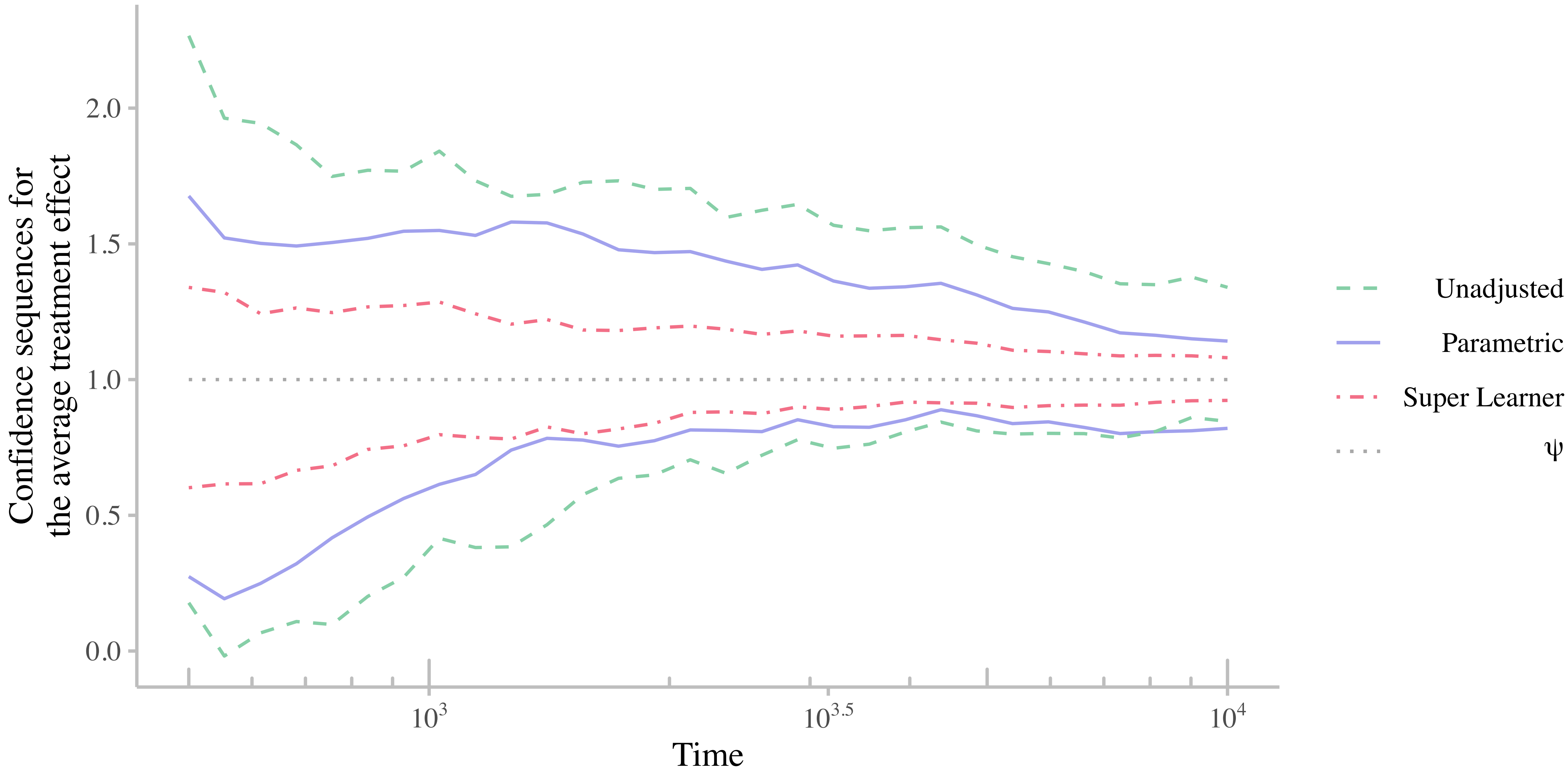
$$\bar{C}_t^{\times} := \left(\widehat{\psi}_t^{\times} \pm \sqrt{t^{-2}(2t\widehat{\sigma}_t^2 + 1) \cdot \log\left(\alpha^{-1} \sqrt{t\widehat{\sigma}_t^2 + 1} \right)} \right)$$

forms an AsympCS for the ATE ψ.

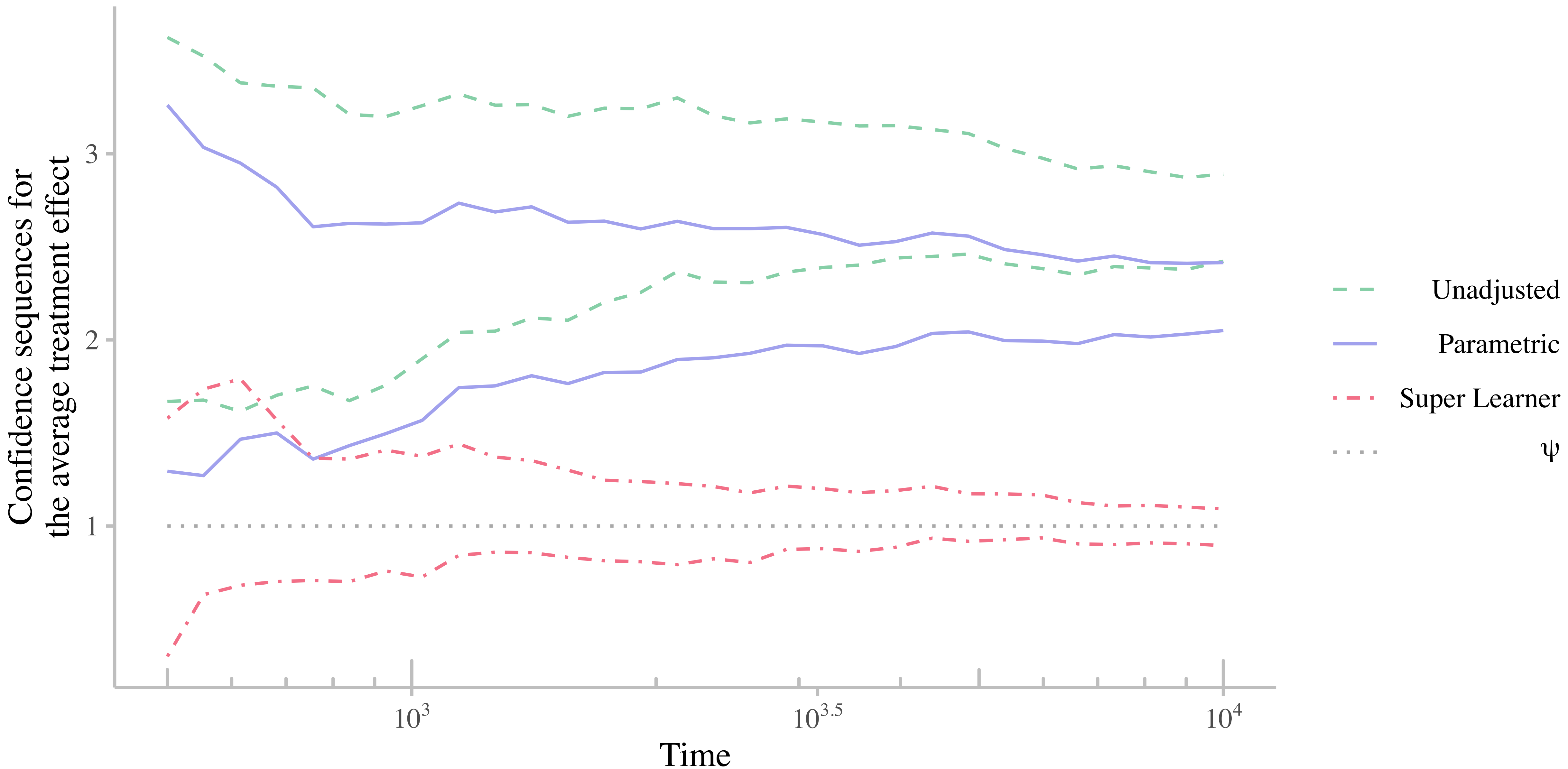
Applies to both randomized expts and observational studies

The usual fixed-n assumption is $o_P(\sqrt{1/t})$, incomparable +ours.

In a randomized experiment, using better regression estimators yields tighter AsympCSs, but all are valid, permitting inference at stopping times.



In observational studies, only consistent regression estimators yield valid AsympCSs (same as fixed-n setting)



Theorem (AsympCS for time-varying treatment effects)

Suppose now that we have the individual treatment effects $\psi_t = \mathbb{E}(Y_t^1 - Y_t^0)$.

Suppose $\frac{1}{t}\sum_{i=1}^{t} \| \widehat{\mu}_t^a(X_i) - \mu^a(X_i) \| \| \widehat{\pi}(X_i) - \pi(X_i) \| = o(\sqrt{\log t / t})$ and

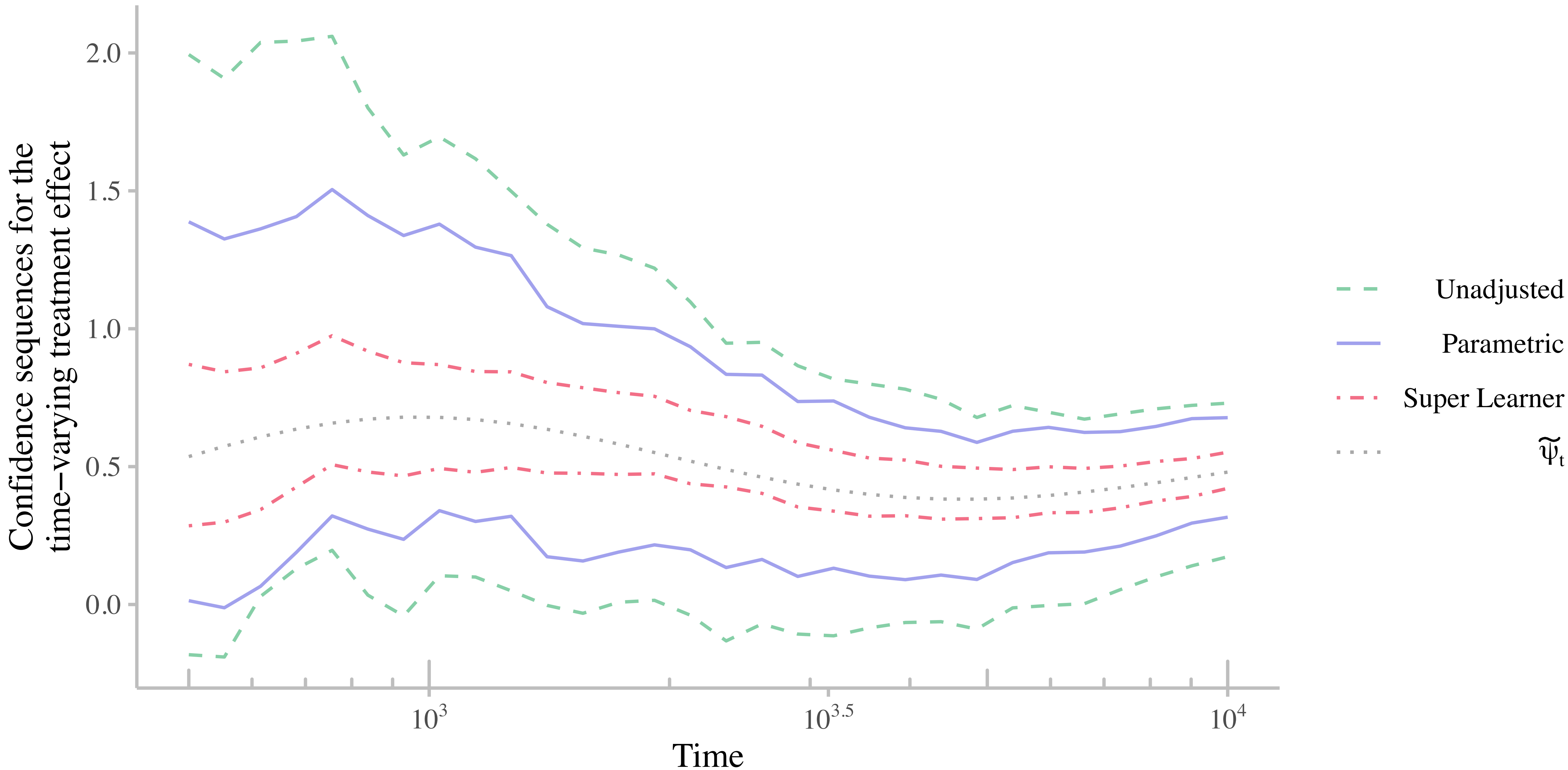
$\sup_i \| \widehat{\mu}_t^a(X_i) - \mu^a(X_i) \| = o(1)$. Then,

$$\widetilde{C}_t^{\times} := \left(\widehat{\psi}_t^{\times} \pm \sqrt{t^{-2}(2t\widehat{\sigma}_t^2 + 1) \cdot \log\left(\alpha^{-1}\sqrt{t\widehat{\sigma}_t^2 + 1} \right)} \right)$$

forms an AsympCS for the *running average of the ITEs* $\widetilde{\psi}_t := \frac{1}{t}\sum_{i=1}^{t} \psi_i$.

If treatment effects are constant over time, $\widetilde{C}_t^{\times}$ captures the ATE!

Our AsympCSs can capture time-varying treatment effects.



The paper has delta method to extend these bounds to asymptotically linear estimators (eg: general semiparametric estimation).

Outline of second half

- ✓ Core definition: confidence sequence
- ✓ A simple, explicit nonparametric example
- ✓ Asymptotic confidence sequences

Summary

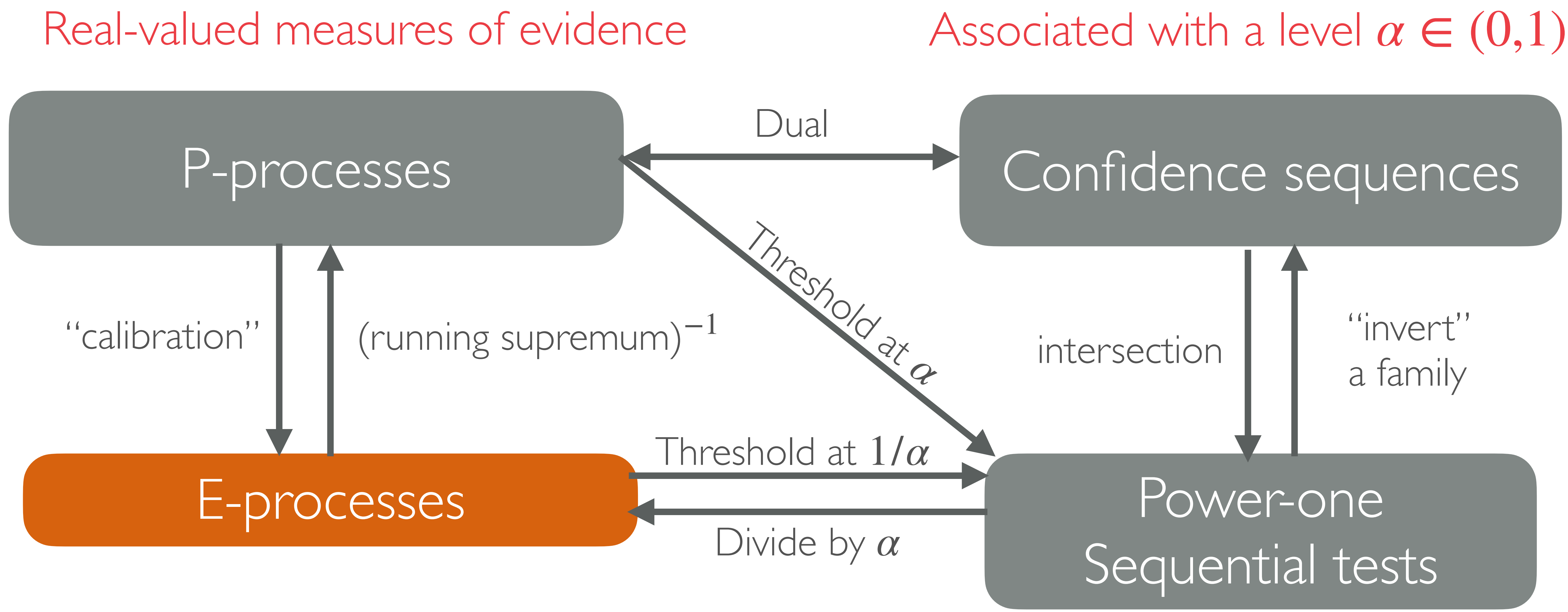
1. Confidence sequences are sequences of confidence intervals that are valid at arbitrary stopping times.

2. Sequential estimation and testing are dual problems. All CSs are obtained by inverting families of sequential tests.

3. Can construct tight CSs even in nonparametric settings.

4. "Time-uniform central limit theory" and asymptotic CSs allow for sequential doubly-robust causal inference in observational settings, and more generally sequential semiparametrics.

Sequential anytime-valid inference (SAVI)



Game-theoretic methods are very practical

1. Election auditing: the state-of-the-art post-election audits are now based on betting for sampling without replacement.

2. A/B testing: our A/B tests are being used by Amazon, Netflix in public-facing software.

3. On and off-policy evaluation: our confidence sequences are deployed at Adobe, Microsoft in public-facing software.

Adobe's Statistical Methodology: Any Time Valid Confidence Sequences

A **Confidence Sequence** is a sequential analog of a **Confidence Interval**, e.g. if you repeat your experiments one hundred times, and calculate an estimate of the mean metric and its associated 95%-Confidence Sequence for every new user that enters the experiment. A 95% Confidence Sequence will include the true value of the metric in 95 out of the 100 experiments that you ran. A 95% Confidence Interval could only be calculated once per experiment in order to give the same 95% coverage guarantee; not with every single new user. Confidence Sequences therefore allow you to continuously monitor experiments, without increasing False Positive error rates.

The difference between confidence sequences and confidence intervals for a single experiment is shown in the animation below:

src: https://experienceleague.adobe.com/docs/journey-optimizer/using/campaigns/content-experiment/experiment-calculations.html

Growthbook is a Y-Combinator startup

GrowthBook's implementation

There are many approaches to sequential testing, several of which are well explained and compared in this Spotify blogpost.

For GrowthBook, we selected a method that would work for the wide variety of experimenters that we serve, while also providing experimenters with a way to tune the approach for their setting. To that end, we implement Asymptotic Confidence Sequences introduced by Waudby-Smith et al. (2023); these are very similar to the Generalized Anytime Valid Inference confidence sequences described by Spotify in the above post and introduced by Howard et al. (2022), although the Waudby-Smith et al. approach more transparently applies to our setting.

src: https://docs.growthbook.io/statistics/sequential

Stuff not covered in the tutorial

1. Multiple hypothesis testing (eg: the e-BH procedure)
2. Sequential changepoint detection and localization using e-processes and CSs (eg: the e-detector)
3. Connections to Bayes, empirical Bayes and PAC-Bayes (eg: prior-posterior ratio martingale, improper priors, compound e-values)
4. Martingale concentration (eg: time-uniform Chernoff bounds)
5. Multivariate CSs (for vectors, matrices, etc.)
6. Universal inference (a simple, general e-value and e-process)
7. Decision making with e-values (eg: post-hoc validity)

Some current and future directions

1. For a new (nonparametric) problem, how do we *design* the game and *learn* to bet?

2. When do *nontrivial* test martingales (not) exist?
   When do *nontrivial* test supermartingales (not) exist?
   When do *nontrivial* e-processes (not) exist?

3. How do we move beyond testing and estimation to, say, other problems in statistics?

4. How do we tie together game-theoretic statistics with game-theoretic probability?

Peter Grunwald

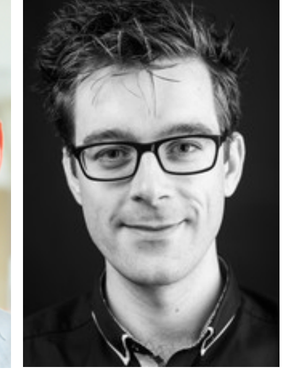
Glenn Shafer

Volodya Vovk

Ruodu Wang

Johannes Ruf

Martin Larsson



Wouter Koolen

Shubhanshu Shekhar

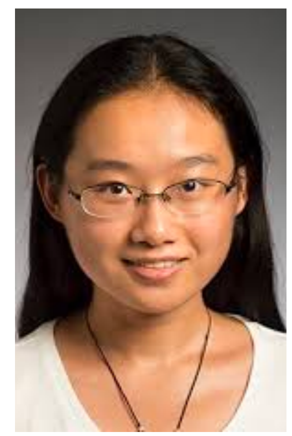
YJ Choe

Steve Howard

Akshay Balsubramani

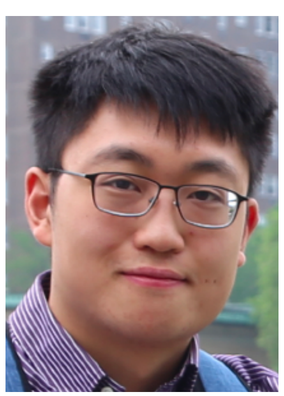
Robin Dunn

Sasha Podkopaev

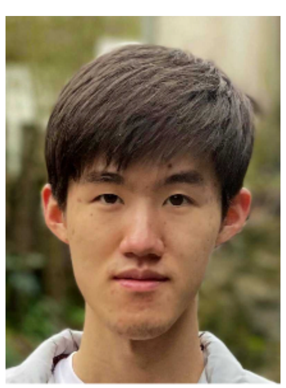


Boyan Duan

Jaehyeok Shin

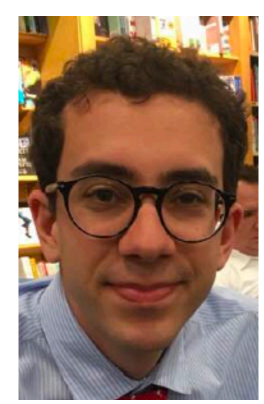


Neil Xu



Hongjian Wang

Ben Chugg



Justin White-house

Ian Waudby-Smith

Tudor Manole

Focus: foundational papers (cutting across problems)

Time-uniform Chernoff bounds via nonnegative supermartingales
(+S. Howard, J. Sekhon, J. McAuliffe), Probability Surveys, 2020

Universal inference *(+L. Wasserman, S. Balakrishnan),* PNAS, 2020

A unified recipe for deriving (time-uniform) PAC-Bayes bounds
(+B. Chugg, H. Wang), JMLR, 2023

Admissible anytime-valid inference must rely on nonnegative martingales
(+M. Larsson, J. Ruf, W. Koolen), arXiv, 2020

The numeraire e-variable and reverse information projection
(+M. Larsson, J. Ruf), Annals of Stat. 2025

Distribution-uniform anytime-valid inference *(+ I. Waudby-Smith, E. Kennedy), arXiv*

The extended Ville's inequality for nonintegrable nonnegative supermartingales
(+H. Wang), Bernoulli, 2025

Randomized & exchangeable improvements of Markov, Chebyshev & Chernoff's inequalities *(+T. Manole), Statistical Science, 2025*

On the existence of powerful p-values and e-values for composite hypotheses
(+Z. Zhang, R. Wang), Annals of Statistics, 2025

A composite generalization of Ville's martingale theorem using e-processes
(+M. Larsson, J. Ruf, W. Koolen), Elec. J of Probability, 2023

Combining evidence across filtrations *(+ YJ. Choe),* arXiv

Positive semidefinite matrix supermartingales *(+ H. Wang),* arXiv

On stopping times of power-one sequential tests: tight lower and upper bounds.
(+ S. Agrawal), arXiv

Focus: testing (specific problems)

Testing exchangeability: fork-convexity, supermartingales and e-processes
(+M. Larsson, J. Ruf, W. Koolen), Intl J of Approx Reasoning. 2025

Nonparametric two-sample testing by betting *(+ S. Shekhar),* IEEE TIT'23

Sequential kernelized independence testing
(+A. Podkopaev, S. Kasivishwanathan, P. Blöbaum), ICML, 2024

Sequential Monte-Carlo testing by betting *(+L. Fischer), JRSSB, 2025*

Comparing sequential forecasters *(+YJ. Choe), Operations Research, 2023*

Huber-robust likelihood ratio tests for composite nulls and alternatives
(+A. Saha), arXiv

Interactive martingale tests for the global null *(+ B. Duan, L. Wasserman), EJS'20*

Nonparametric iterated-logarithm extensions of the sequential generalized LRT
(+J. Shin, A. Rinaldo), IEEE J Selected Areas in IT, 2021

Sequential Kernelized Stein Discrepancy *(+D. Martinez-Taboada), AISTATS'25*

Anytime-valid t-tests & CSs for Gaussian means with unknown variance
(+H. Wang), Sequential Analysis, 2025

Sequential predictive two-sample & independence testing
(+ A. Podkopaev), NeurIPS'23

E-variables for hypotheses generated by constraints *(+ M. Larsson, J. Ruf), arXiv*

Improving Wald's (approximate) SPRT by avoiding overshoot *(+ L. Fischer), arXiv*

Anytime-valid inference for double/debiased machine learning of causal parameters
(+ A. Dalal, P. Blöbaum, S. Kasivishwanathan), arXiv

Focus: multiple hypothesis testing

False discovery rate control with e-values *(+ R. Wang),* JRSSB, 2022

Post-selection inference for e-value based confidence intervals
(+Z. Xu, R. Wang), EJS'24

E-values as unnormalized weights in multiple testing
(+N. Ignatiadis, R. Wang), Biometrika, 2023

A unified framework for bandit multiple testing *(+Z. Xu, R. Wang), NeurIPS'21*

Online multiple testing with e-values *(+Z. Xu), AISTATS, 2024*

An online generalization of the (e-)BH procedure *(+L. Fischer, Z. Xu), arXiv*

Asymptotic & compound e-values: multiple testing & empirical Bayes.
(+N. Ignatiadis, R. Wang), arXiv

Bringing closure to FDR control: beating the e-BH procedure
(+Z. Xu, R. de Heide, L. Fischer, A. Solari, J. Goeman), arXiv

Active multiple testing with proxy p-values and e-values
(+Z. Xu, C. Wang, L. Wasserman, K. Roeder), arXiv

Admissible online closed testing must employ e-values *(+L. Fischer), arXiv*

Anytime-valid FDR control with the stopped e-BH procedure
(+H. Wang, S. Dandapanthula), arXiv

More powerful multiple testing under dependence via randomization *(+Z. Xu), arXiv*

Multiple testing with anytime-valid Monte-Carlo p-values *(+L. Fischer, T. Barry), arXiv*

Merging uncertainty sets via majority vote. *(+M. Gasparin), arXiv*

Focus: estimation (confidence sequences)

Time-uniform, nonparametric, nonasymptotic confidence sequences
(+S. Howard, J. Sekhon, J. McAuliffe), The Annals of Stat., 2021

Estimating means of bounded random variables by betting
(+I. Waudby-Smith), J. Royal Stat Society B, 2023 (Discussion paper)

Time-uniform central limit theory and asymptotic confidence sequences
(+I. Waudby-Smith, D. Arbour, R. Sinha, E. Kennedy), Annals of Stat., 2024

Martingale methods for sequential estimation of convex functionals & divergences
(+T. Manole), IEEE Trans. Info. Theory, 2023

Off-policy confidence sequences *(+ N. Karampatziakis, P. Mineiro), ICML'21*

Anytime-valid off-policy inference in contextual bandits
(+I. Waudby-Smith, L. Wu, N. Karampatziakis, P. Mineiro), ACM/IMS J. Data Sci.'24

Catoni-style confidence sequences for heavy-tailed mean estimation
(+H. Wang), Stochastic Proc. & Applications, 2023

Sequential estimation of quantiles with applications to A/B-testing & bandits
(+S. Howard), Bernoulli, 2022

Huber-robust confidence sequences *(+H. Wang), AISTATS, 2023*

Confidence sequences for sampling without replacement
(+I. Waudby-Smith), NeurIPS, 2020

Sharp empirical Bernstein bounds for the variance of bounded random variables
(+D. Martinez-Taboada), arXiv

On the near-optimality of betting confidence sets for bounded means
(+S. Shekhar), arXiv

Focus: estimation (vector or matrix CSs)

Time-uniform, nonparametric, nonasymptotic confidence sequences
(+S. Howard, J. Sekhon, J. McAuliffe), The Annals of Stat., 2021

Time-uniform central limit theory and asymptotic confidence sequences
(+I. Waudby-Smith, D. Arbour, R. Sinha, E. Kennedy), Annals of Stat., 2024

Time-uniform confidence spheres for means of random vectors
(+B. Chugg, H. Wang), TMLR, 2025

Mean estimation in Banach spaces under infinite variance & martingale dependence
(+J. Whitehouse, B. Chugg, D. Martinez-Taboada), arXiv

Sharp matrix empirical Bernstein inequalities *(+H. Wang), arXiv*

Empirical Bernstein in smooth Banach spaces *(+D. Martinez-Taboada), arXiv*

Time-uniform self-normalized concentration for vector-valued processes
(+J. Whitehouse, S. Wu), arXiv

Focus: changepoint analysis

E-detectors: a nonparametric framework for online changepoint detection
(+J. Shin, A. Rinaldo), New England J of Statistics & Data Science, 2023

Reducing sequential change detection to sequential estimation
(+S. Shekhar), ICML'24

Sequential change detection via backward confidence sequences
(+S. Shekhar), ICML'23

Multiple testing in multi-stream sequential change detection
(+S. Dandapanthula), arXiv

Post-detection inference for sequential changepoint localization *(+A. Saha), arXiv*

Focus: auditing

Auditing fairness by betting
(+B. Chugg, S Cortes-Gomez, B. Wilder), NeurIPS, 2023

RiLACS: risk limiting (election) audits via confidence sequences
(+I. Waudby-Smith, P. Stark), EVoteID (Best paper), 2021

Risk-limiting financial audits via weighted sampling without replacement
(+S. Shekhar, Z. Xu, Z. Lipton, P. Liang), UAI, 2023

Sequentially auditing differential privacy
(+T. Gonzalez-Lara, M. Dulce-Rubio, M. Ribero), in submission

Foundational (recent) papers by other authors

Testing by betting (G. Shafer), *JRSSA'20* (Discussion paper)

Safe testing (P. Grünwald, R. de Heide, W. Koolen), *JRSSB'24* (Discussion paper)

E-values: Calibration, combination & applications (V. Vovk and R. Wang), *AoS, 2021*

Beyond Neyman-Pearson (P. Grünwald), *PNAS'25*

The e-posterior (P. Grünwald), *Phil. Trans. Royal Society'25*

Reverse information projections & optimal e-statistics
(T. Lardy, P. Grünwald, P. Harremoes), *IEEE TIT'24*

Tight CSs & the regret of universal portfolio (F. Orabona, KS. Jun), *IEEE TIT'24*

+many old papers by Robbins, Cover, Lai, Siegmund, Vovk, etc.

Surveys and books

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

Game-theoretic foundations for probability and finance (G. Shafer, V. Vovk), *Wiley 2019*

Probability and finance: it's only a game (G. Shafer, V. Vovk), *Wiley 2001*

Likelihood, replicability and Robbins' confidence sequences
(L. Pace, A. Salvan), *Intl. Stat. Review, 2021*

Hypothesis testing with e-values *(+R. Wang),* Foundations & Trends in Statistics'25

Game-theoretic statistics and safe anytime-valid inference
(+P. Grunwald, V. Vovk, G. Shafer), Statistical Science, 2023