A martingale theory of evidence

Aaditya Ramdas

Dept. of Statistics and Data Science (75%)
Machine Learning Dept. (25%)
Carnegie Mellon University

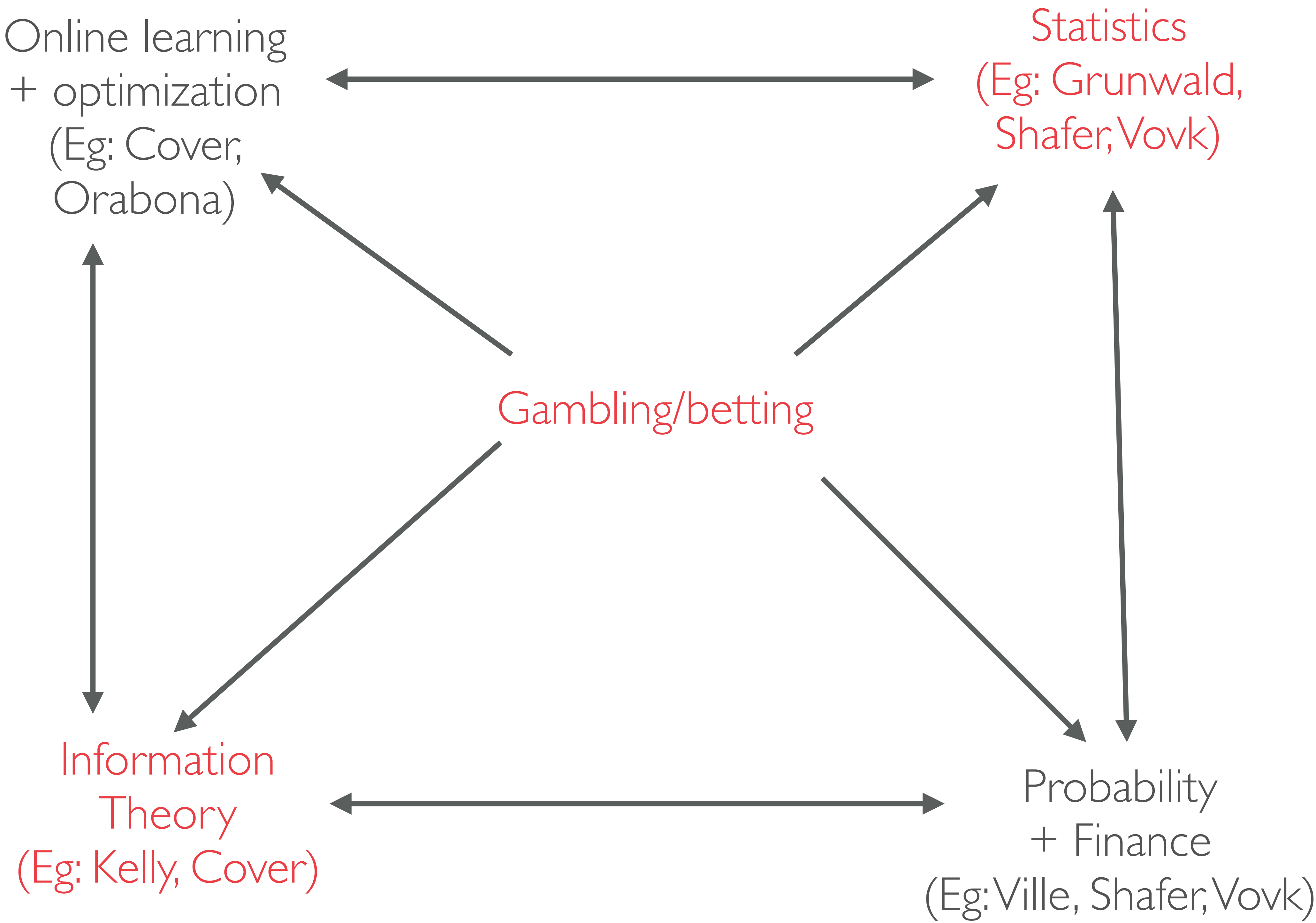
Amazon Research (20%)

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

Outline of this lecture series

- Yesterday: game-theoretic testing
2. Now: game-theoretic estimation
3. Today afternoon: game-theoretic change detection

Quick recap of game-theoretic testing


Online learning
+ optimization
(Eg: Cover,
Orabona)
Statistics
(Eg: Grunwald,
Shafer, Vovk)
Gambling/betting
Information
Theory
(Eg: Kelly, Cover)
Probability
+ Finance
(Eg: Ville, Shafer, Vovk)

Core idea: Testing by betting

In order to test a hypothesis, one sets up a game such that:
if the null is true, no strategy can systematically make (toy) money,
but if the null is false, then a good betting strategy can make money.

Wealth in the game is directly a measure of evidence against the null.

Each strategy of the statistician = a different estimator or test statistic.
So there are "good" and "bad" strategies for betting,
just as there are good and bad estimators or test statistics.

Testing and estimation == game and strategy design.

A p-process (or anytime-valid p-value) for a null $H_0 : P \in \mathscr{P}$ is a sequence $(p_t)_{t \geq 1}$ that satisfies

For any stopping time $\tau, P \in \mathscr{P} : P(p_\tau \leq \alpha) \leq \alpha$.

Johari et al. (2015, 2021),
Howard, Ramdas, et al. (2018, 2021)

An e-value for H_0 is a $[0,\infty]$-valued r.v. e s.t. $\forall P \in \mathscr{P}, \mathbb{E}_P(e) \leq 1$. (e for evidence or expectation)

An e-process for H_0 is a sequence of e-values $(e_t)_{t \geq 1}$

$$\sup_{\tau} \sup_{P \in \mathscr{P}} \mathbb{E}_P(e_\tau) \leq 1.$$

Eg: nonnegative martingales, supermartingales and more.

Howard, Ramdas, et al. (2018-2021)
Grunwald et al. (2019-2021)
Shafer (2020), Vovk & Wang (2021)

Summary

Testing by betting is a simple framework for hypothesis testing that yields sequential, anytime-valid inference.

Optimal gambling strategies are based on likelihood ratios.
Composite alternatives are handled using mixtures (hedging).
Composite nulls are handled using reverse information projections, or via universal inference (maximum-likelihood under the null).

(Composite) Nonnegative (super)martingales are secretly likelihood ratios, even when no reference measure exists.

E-processes exist more generally, even when nonnegative supermartingales do not exist. They are central objects: necessary and sufficient for sequential testing.

So what about estimation?

A nontrivial nonparametric example

Estimating means of bounded random variables by betting

(J Royal Stat Society B, 2023, discussion paper)

Ian Waudby-Smith

Setting 1

Note: rich nonparametric set of distributions without a reference measure (hence no likelihood ratios)

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

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

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

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

A3: Betting — significantly tighter!

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

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.

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

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

(Another motivation: (L_n, U_n) should not contradict (L_m, U_m) for any $m > n$. With pointwise CIs, intersection $= \emptyset$ a.s., but with 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:

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

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

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.

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 $K_t^{(m)} = K_{t-1}^{(m)} \cdot (1 + \lambda_t^{(m)}(X_t - m))$

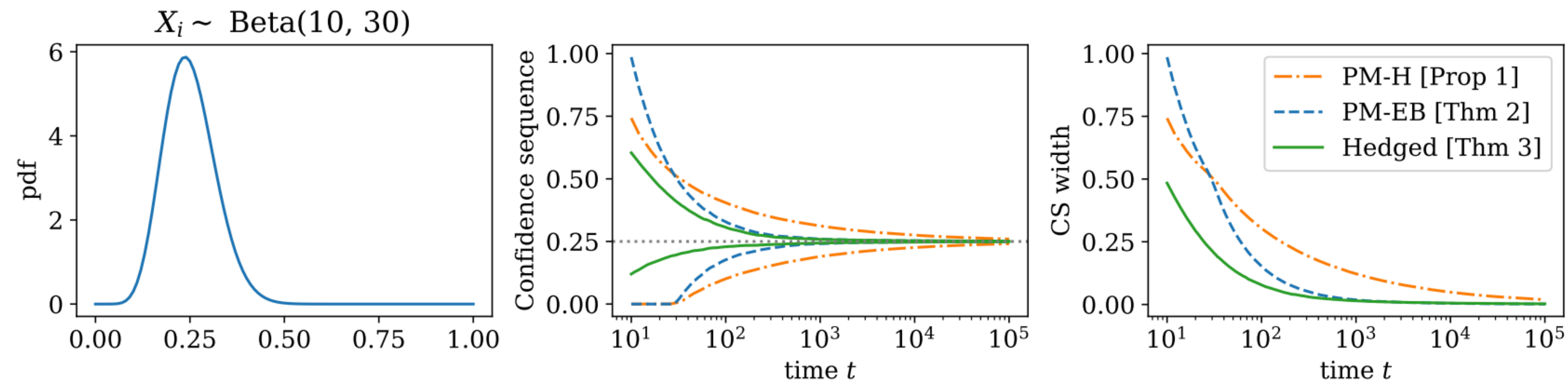
$$C_t := \left\{m \in [0,1] : K_t^{(m)} < 1/\alpha\right\}$$

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

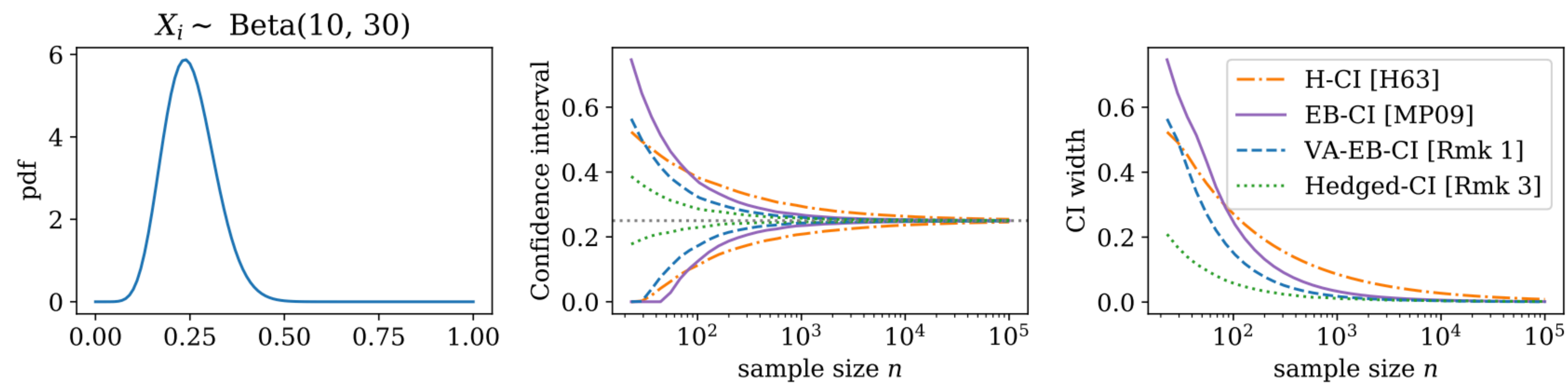
Theorem: For any betting strategy, $(C_t)_{t\geq 1}$ is a confidence sequence for the true mean μ.

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

Time-uniform confidence sequences



Fixed-time confidence intervals



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

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

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

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

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

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

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

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

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

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

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

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

But, how should we bet? (Option 1: GRAPA)

Growth Rate Adaptive to the Particular Alternative

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

But we don't know P. Approximate solution:
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)

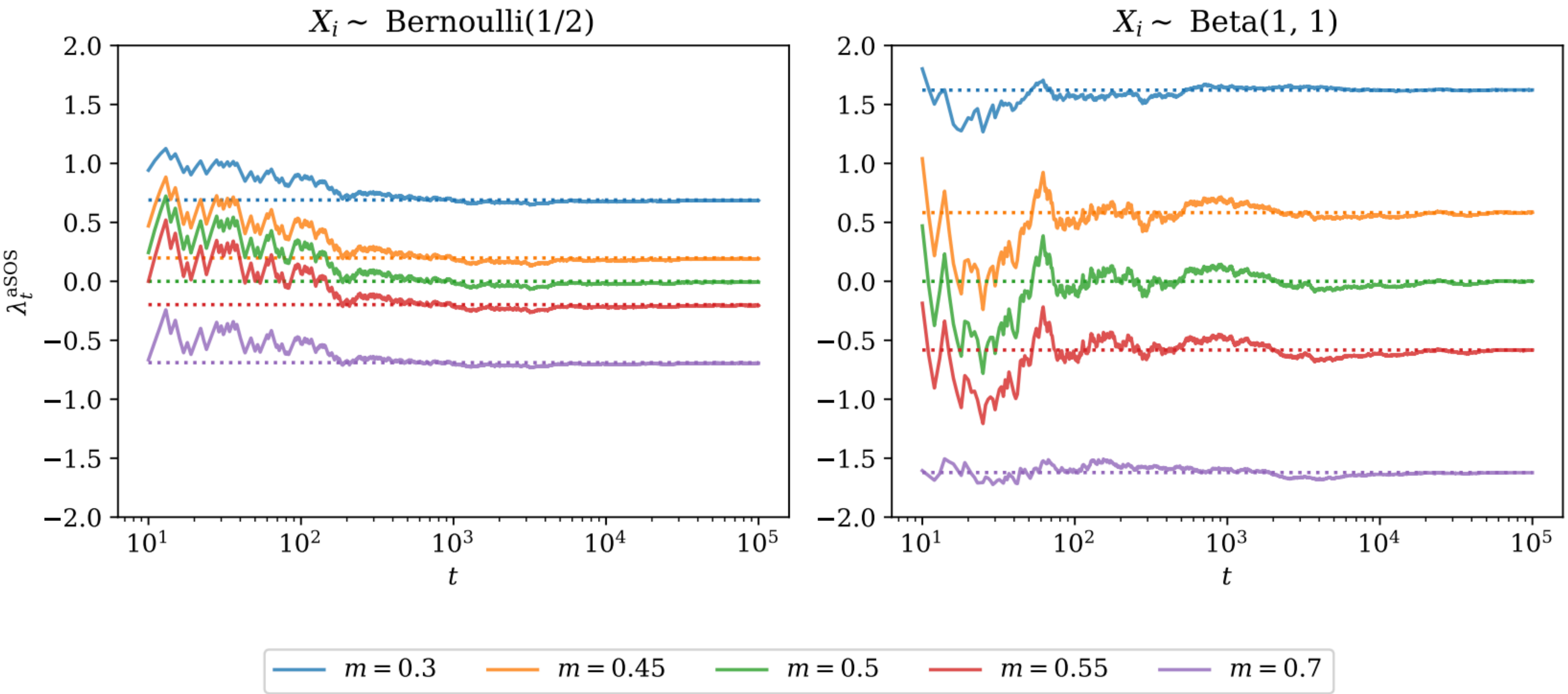


Figure 7: λ_t^{aSOS} for various values of m under two distributions: Bernoulli(1/2) and Beta(1, 1). The dotted lines show the 'oracle' bets, meaning λ_t^{aSOS} with estimates of the mean and variance replaced by their true values. As time passes, bets stabilize and approach their oracle quantities.

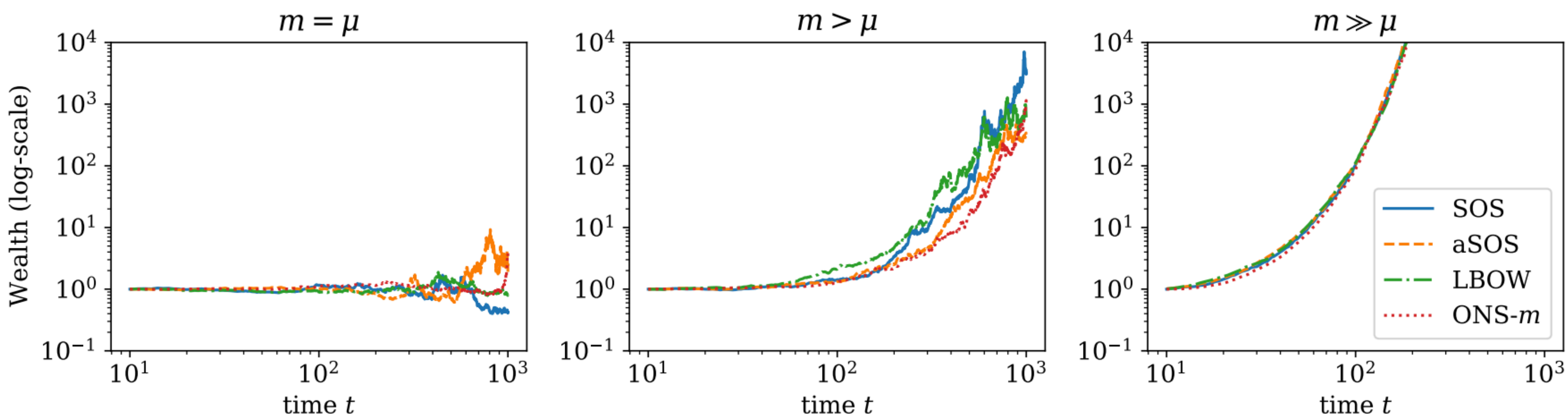
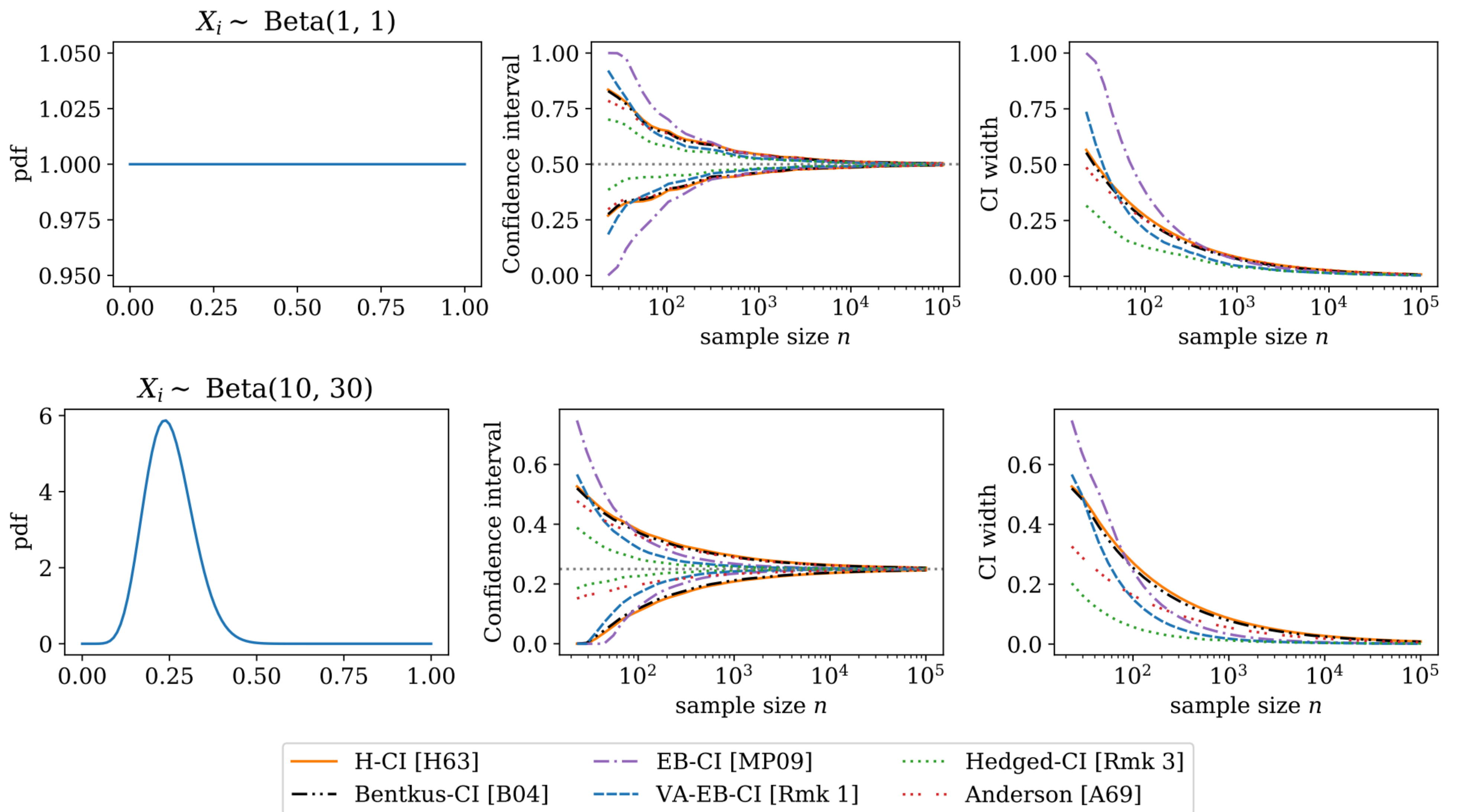


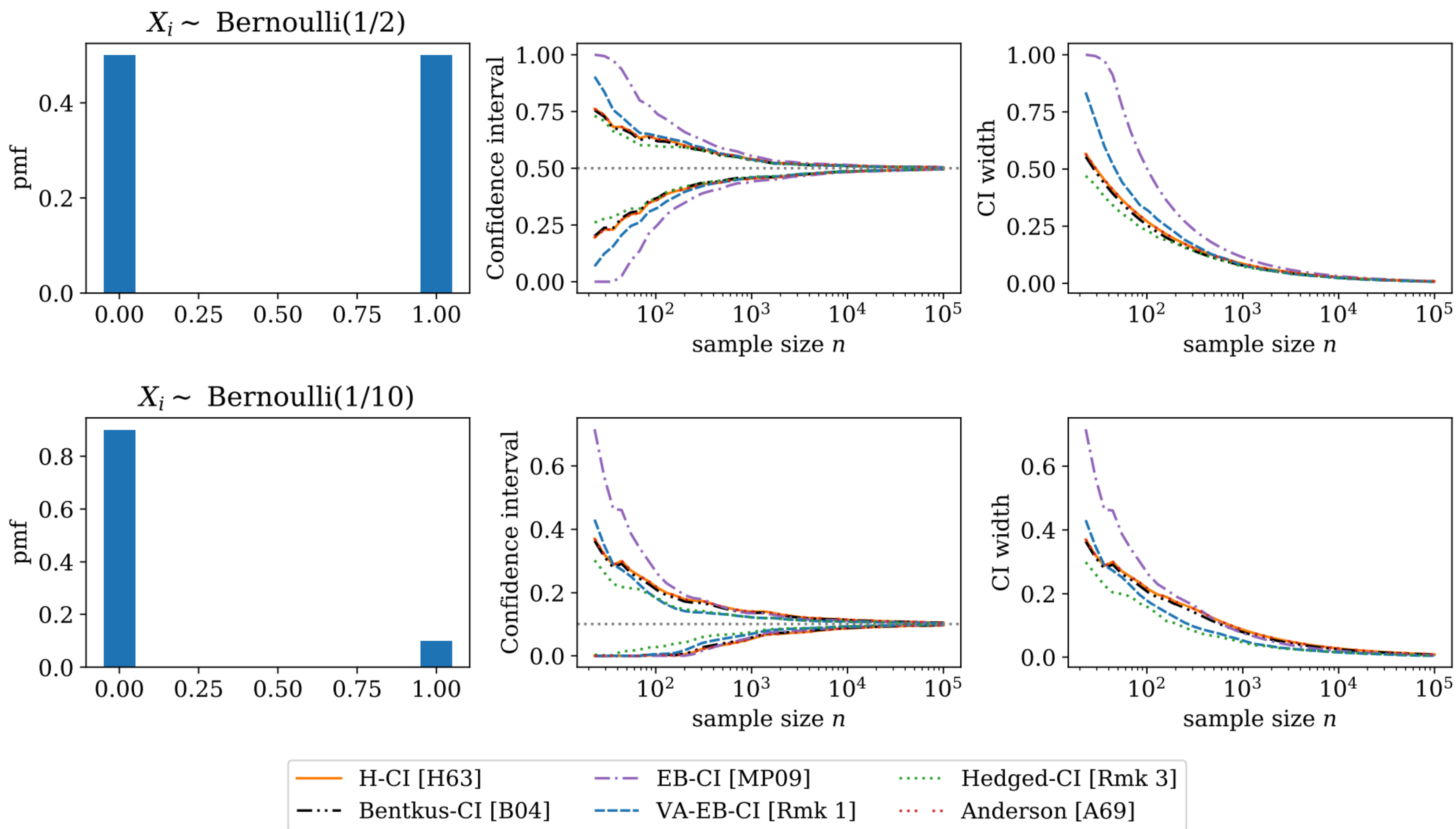
Figure 8: Comparison of the wealth process under various game-theoretic betting strategies with 100 repeats. In this example, the 1000 observations are drawn from a Beta(10, 10) distribution, and the candidate means m being tested are 0.5, 0.51, and 0.55 (from left to right). Notice that these strategies perform similarly, but have varying computational costs (see Table 2).



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)



$X_i \sim$ Bernoulli(1/2)
$X_i \sim$ Bernoulli(1/10)
pmf
Confidence interval
CI width
sample size n
H-CI [H63]
Bentkus-CI [B04]
EB-CI [MP09]
VA-EB-CI [Rmk 1]
Hedged-CI [Rmk 3]
Anderson [A69]

Games, capital processes are intrinsic to (many/most/all?) testing problems

$$K_t^{(\mu)} := \prod_{i \le t} (1 + \lambda_i (X_i - \mu))$$

$K_t^{(\mu)}$ is a test martingale
if and only if
every X_i has conditional mean μ.

Thus, the capital process being a nonnegative martingale
— which is the only property we used for validity —
is not just an implication of the problem statement,
it is actually logically equivalent to the problem statement.

This is not just true for the presented problem, but a slew of other nonparametric problems like independence testing, heavy-tailed estimation, etc. We can prove that SAVI inference MUST be based on capital processes.

Ramdas, Ruf, Larsson, Koolen
(arXiv:2009.03167)

A multi-round game against an adaptive, constrained adversary

Adversary first picks $\mu \in [0,1]$.

At each time t

1. Statistician discloses bets for every m, depending on past.

2. Adversary then picks a distribution $Q_t^\mu \in \mathcal{Q}^\mu$, which could also depend on the past, and on the bets.

3. Nature verifies that rules are being followed, draws $X_t \sim Q_t$ and presents it to the statistician.

$\mathcal{P}^\mu$ is a "closed, fork-convex" set. Sequential analog of convexity.

Setting 2

$x_1, \ldots, x_N$ are fixed non-random numbers in $[0,1]$.

$X_t \mid \{X_1, \ldots, X_{t-1}\} \sim \text{Unif}[\{x_1, \ldots, x_N\} \backslash \{X_1, \ldots, X_{t-1}\}]$.

Q1. How can we construct a confidence interval for $\mu := \sum_{i=1}^{N} x_i / N$?

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

See the paper for details

A2. Serfling (1970s), Bardenet-Maillard (2014), etc.

A3: Betting — significantly tighter!

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

Time-uniform, nonparametric, nonasymptotic confidence sequences

+ "Time-uniform Chernoff bounds"

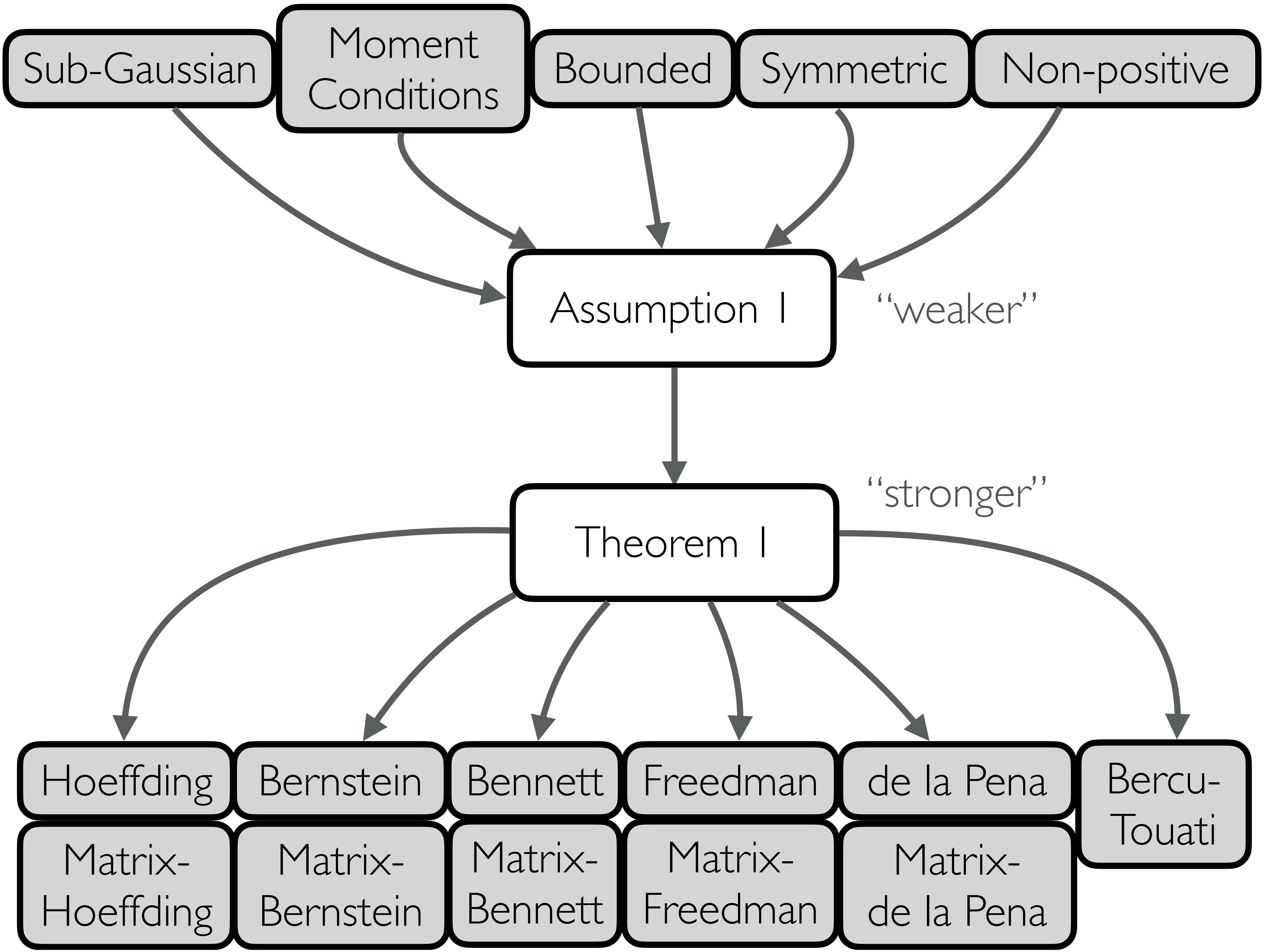



Steve Howard

Jas Sekhon

Jon McAuliffe

(AoS'21, Probability Surveys'20)


Sub-Gaussian
Moment Conditions
Bounded
Symmetric
Non-positive
Assumption 1
"weaker"
"stronger"
Theorem 1
Hoeffding
Bernstein
Bennett
Freedman
de la Pena
Bercu-Touati
Matrix-Hoeffding
Matrix-Bernstein
Matrix-Bennett
Matrix-Freedman
Matrix-de la Pena

“‘stronger’”?

(A) Assumptions

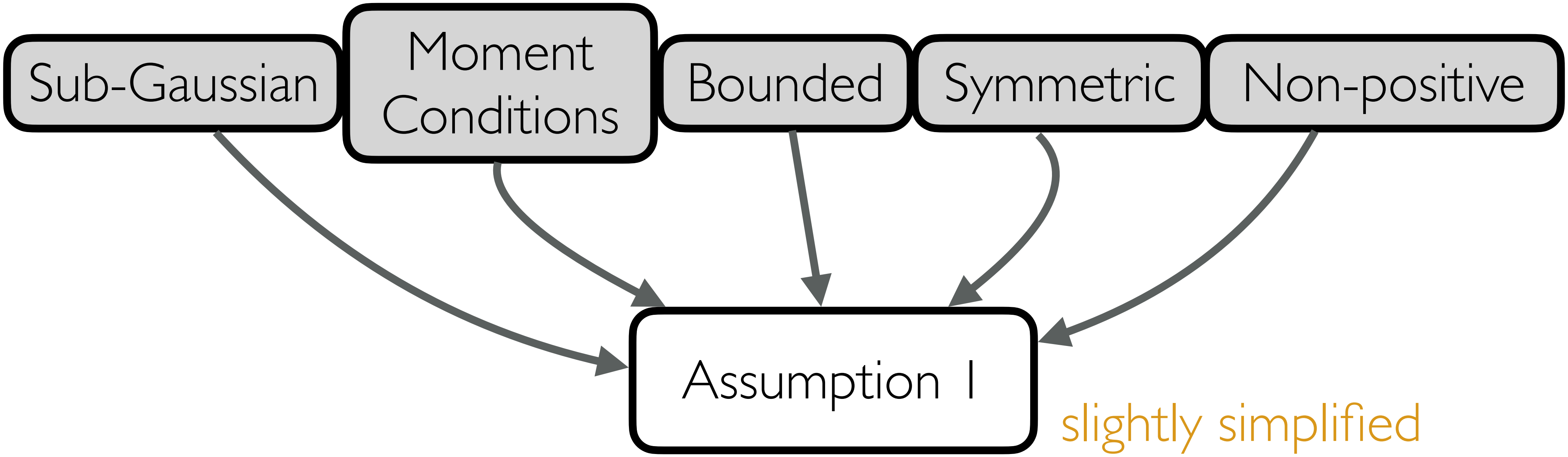
(B) Boundary

(C) Continuous

(D) Dimension

(E) Exponent

Existing result	Our result	[A]	[B]	[C]	[D]	[E]
Bernstein (1927)	Corollary 1(c)		✓	✓	✓	
Bennett (1962, eq. (8b))	Corollary 1(b)	✓	✓	✓	✓	
Hoeffding (1963, Theorem 2)	Corollary 1(a)	✓	✓		✓	
Freedman (1975, Theorem 1.6)	Corollary 1(b)		✓	✓	✓	
Shorack and Wellner (1986, App. B, Ineq. 1)	Corollary 11(b)		✓			
Pinelis (1994, Theorems 3.4, 3.5)	Corollary 10		✓			
van de Geer (1995, Lemma 2.2)	Corollary 11(c)		✓		✓	
Blackwell (1997, Theorem 1)	Corollary 4(a)	✓		✓	✓	
Blackwell (1997, Theorem 2)	Corollary 5				✓	
Blackwell (1997, Theorem 2)	Corollary 4(b)	✓		✓	✓	
de la Peña (1999, Theorems 1.2B, 6.1)	Corollary 6		✓	✓	✓	
de la Peña (1999, Theorem 6.2)	Corollary 7			✓	✓	✓
Bercu and Touati (2008, Theorem 2.1)	Corollary 8		✓		✓	✓
Delyon (2009, Theorem 4)	Corollary 8		✓		✓	
Khan (2009, Theorem 4.2)	Theorem 1(b)		✓	✓	✓	
Khan (2009, Theorem 4.3)	Theorem 1(d)			✓	✓	✓
Tropp (2011, Theorem 1.2)	Corollary 1(b)		✓			
Tropp (2012, Theorem 1.3)	Corollary 1(a)		✓			✓
Tropp (2012, Theorem 1.4)	Corollary 1(c)		✓			
Mackey et al. (2014, Corollary 4.2)	Corollary 1(a)	✓	✓			



Given a "sum process" (S_n), assume that we can find a "variance process" (V_n) and a function $\psi(\lambda)$ such that $\exp(\lambda S_n - \psi(\lambda) V_n)$ is upper bounded by a test supermartingale for any $\lambda \in [0, \lambda_{\max})$.

"Sub-ψ supermartingale or e-process"

Often, $\psi(\lambda)$ is a CGF (log-MGF), $\psi(\lambda) \approx \lambda^2/2$ as $\lambda \to 0$.

Example: subGaussian observations

$\mathbb{E}[\exp(\lambda X_i) \mid \mathcal{F}_{i-1}] \leq \exp(\lambda^2 \sigma_i^2 / 2)$

Define $M_n := \prod_{i=1}^{n} \exp(\lambda X_i - \frac{\lambda^2}{2} \sigma_i^2)$.

If $\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] \leq 0$ and X_i is σ_i-subGaussian, then M_n is an NSM for $\lambda \geq 0$.

Robbins et al. 1960s/70s

Example: observations bounded on one side

Denote $S_n := \sum_{i=1}^{n} X_i$, $V_n = \sum_{i=1}^{n} X_i^2$.

Define $M_n := \exp(\lambda S_n - (-\log(1-\lambda) - \lambda))V_n)$

$$= \prod_{i=1}^{n} \exp(\lambda X_i - \psi_E(\lambda) X_i^2)$$

If $X_i \geq -1$, and $\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] \leq 0$, then M_n is an NSM for $\lambda \in [0,1]$.

Fan et al. (2015)

Assumption 1

$$\exp\{\lambda S_n - \psi(\lambda) V_n\} \le L_n(\lambda), \quad a.s. \quad \forall n \ge 1.$$

$$\psi^\star(u) := \sup_{\lambda \in \mathbb{R}} [\lambda u - \psi(\lambda)] \quad \text{(the Legendre-Fenchel transform)},$$

$$\mathfrak{s}(u) := \frac{\psi(\psi^{\star\prime}(u))}{\psi^{\star\prime}(u)} \quad \text{(the "slope" transform)}.$$

Theorem 1

Suppose (S_n), (V_n), and $\psi(\lambda)$ satisfy Assumption 1, where ψ is strictly convex and twice continuously differentiable with $\psi(0) = \psi'(0) = 0$ and $\sup_\lambda \psi'(\lambda) = \infty$.

Then for any $x, m > 0$, we have

$$\mathbb{P}\left(\exists n \ge 1 : S_n \ge x + \mathfrak{s}\left(\frac{x}{m}\right)(V_n - m)\right) \le (\mathbb{E}L_0) \exp\left\{-m\psi^\star\left(\frac{x}{m}\right)\right\}$$

Eg: Suppose $|X_n| \le b$ and $\mathbb{E}_{n-1} X_n = 0$ for all n,
and let $V_n = \sum_{i=1}^n \operatorname{Var}_{i-1} X_i$.

(Bennett '62)

$$\mathbb{P}(S_m \ge x) \le \exp\left\{-\frac{V_m}{b^2}\mathfrak{h}\left(\frac{bx}{V_m}\right)\right\}$$

$$\mathfrak{h}(u) = (1+u)\log(1+u) - u.$$

(Freedman '75) $\mathbb{P}(\exists n \ge 1 : S_n \ge x \text{ and } V_n \le m) \le \exp\left\{-\frac{m}{b^2}\mathfrak{h}\left(\frac{bx}{m}\right)\right\}$

are special cases of Theorem 1

$$\mathbb{P}\left(\exists n \ge 1 : S_n \ge x + \mathfrak{s}\left(\frac{x}{m}\right)(V_n - m)\right) \le (\mathbb{E}L_0)\exp\left\{-\frac{m}{b^2}\mathfrak{h}\left(\frac{bx}{m}\right)\right\}$$

$$\mathfrak{s}(u) = \frac{1}{b}\left(\frac{bu}{\log(1+bu)} - 1\right)$$

Eg: (bounded max-eigenvalue)

Suppose $\gamma_{\max}(X_n) \leq b$, recall $V_n := \gamma_{\max}(\sum_{i=1}^n \mathrm{Var}_{i-1} X_i)$

Theorem 1

$$\mathbb{P}\left(\exists n \geq 1 : S_n \geq x + \mathfrak{s}\left(\frac{x}{m}\right)(V_n - m)\right) \leq d \exp\left\{-\frac{m}{b^2}\mathfrak{h}\left(\frac{bx}{m}\right)\right\}$$

$$\leq d \exp\left\{-\frac{x^2}{2(m + bx/3)}\right\}$$

strengthens Tropp's matrix-Freedman/Bernstein

Eg: (matrix sub-Gaussian)

Suppose $\mathbb{E}_{n-1} e^{\lambda X_n} \preceq e^{\lambda^2 \sigma_n^2 / 2}$

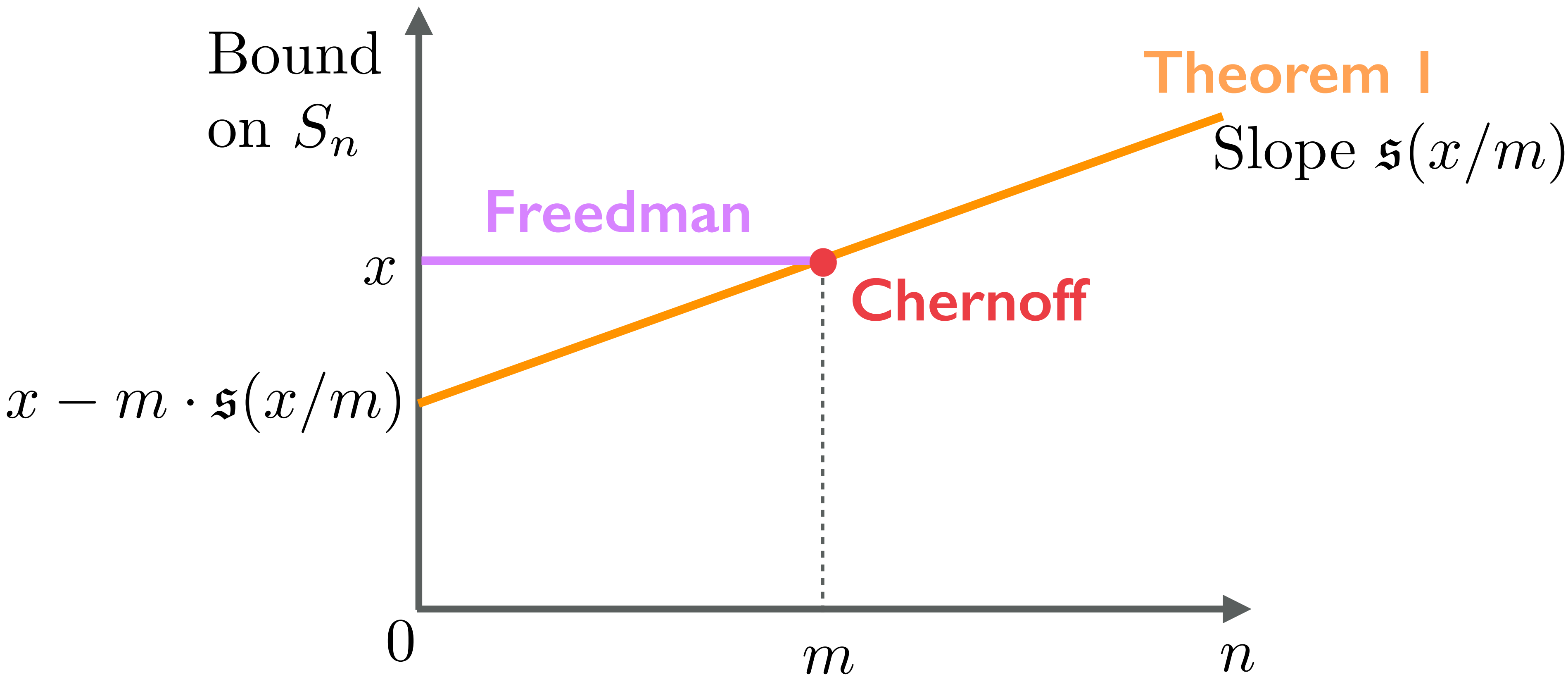
Theorem 1

$$\mathbb{P}\left(\exists n \geq 1 : S_n \geq x + \frac{x}{2m}\left[\gamma_{\max}\left(\sum_{i=1}^{n} \sigma_i^2\right) - m\right]\right) \leq d \exp\left\{-\frac{x^2}{2m}\right\}$$

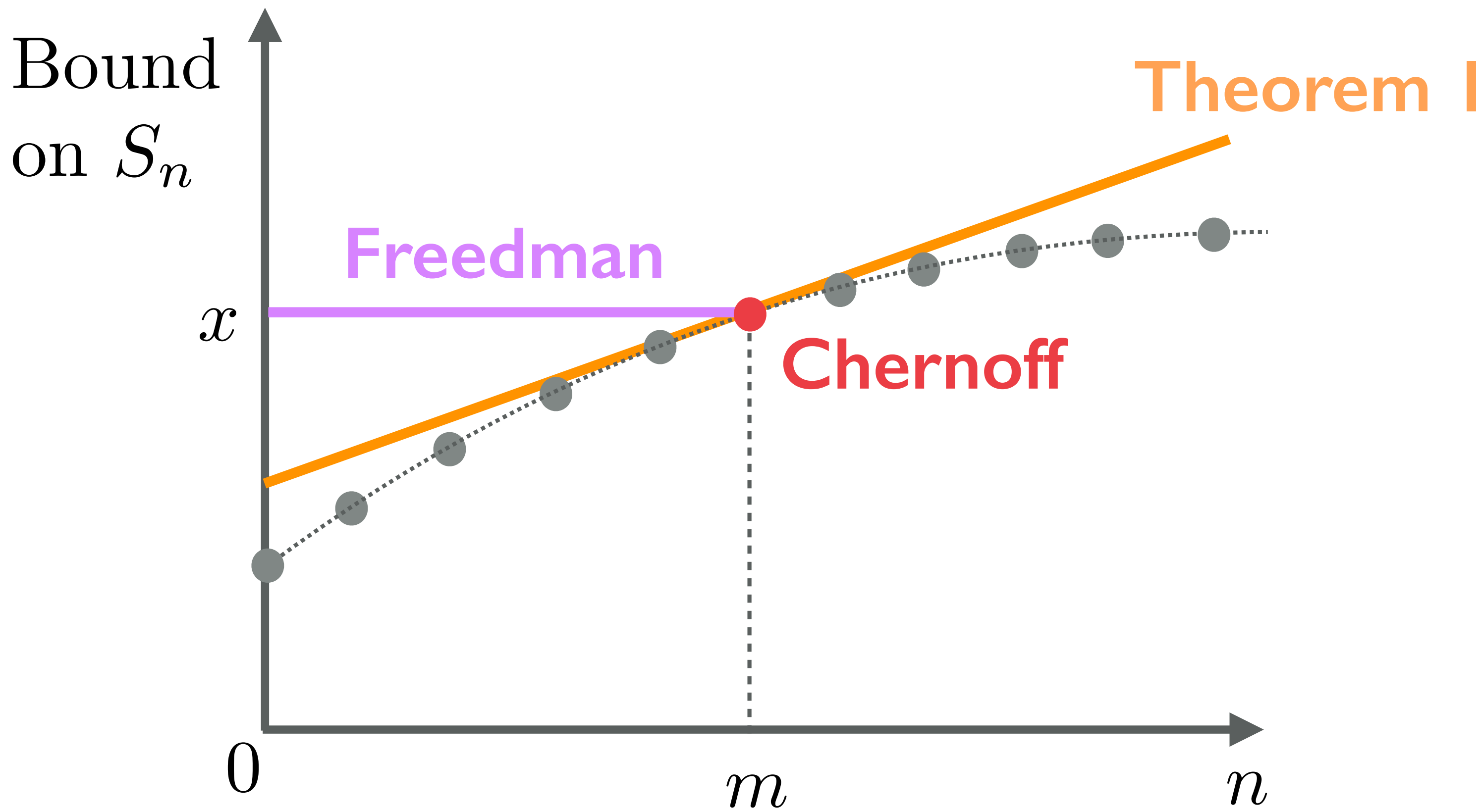
strengthens matrix-Hoeffding in Tropp (2012), Wainwright (2018), Ahlswede-Winter (2002), etc.

Suppose $V_n = n$

Exceedence probability $\alpha = \exp\{-m\psi^*(x/m)\}$.



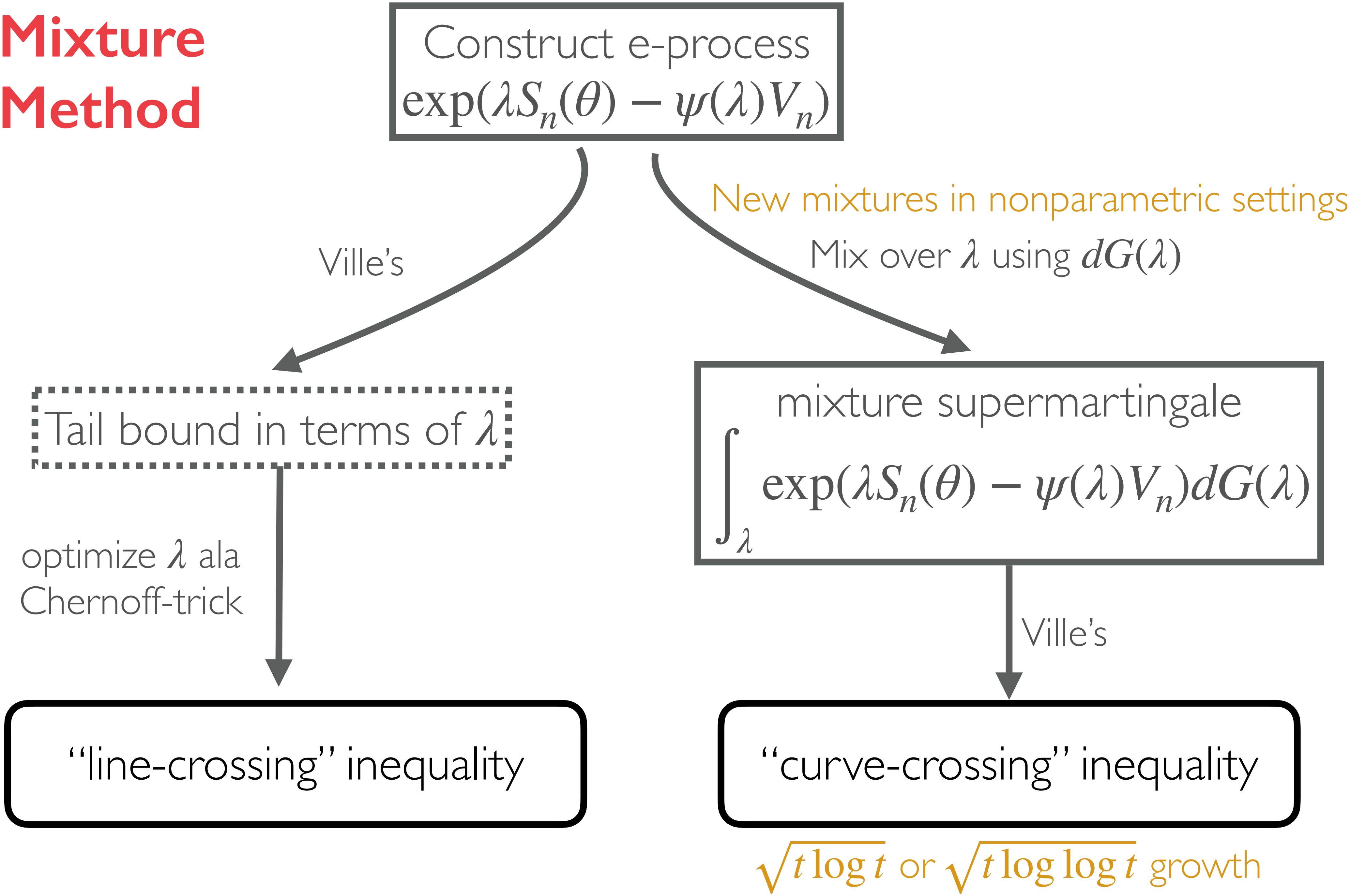
Theorem 1	$\mathbb{P}\left(\exists n \geq 1 : S_n \geq x + (n-m) \cdot \mathfrak{s}(x/m)\right) \leq \alpha$
Freedman	$\mathbb{P}\left(\exists n \geq 1 : S_n \geq x \text{ and } n \leq m\right) \leq \alpha$
Chernoff	$\mathbb{P}(S_m \geq x) \leq \alpha$



Underlying every Chernoff bound is a uniform bound.

Our uniform bound is tangent to the pointwise curve.

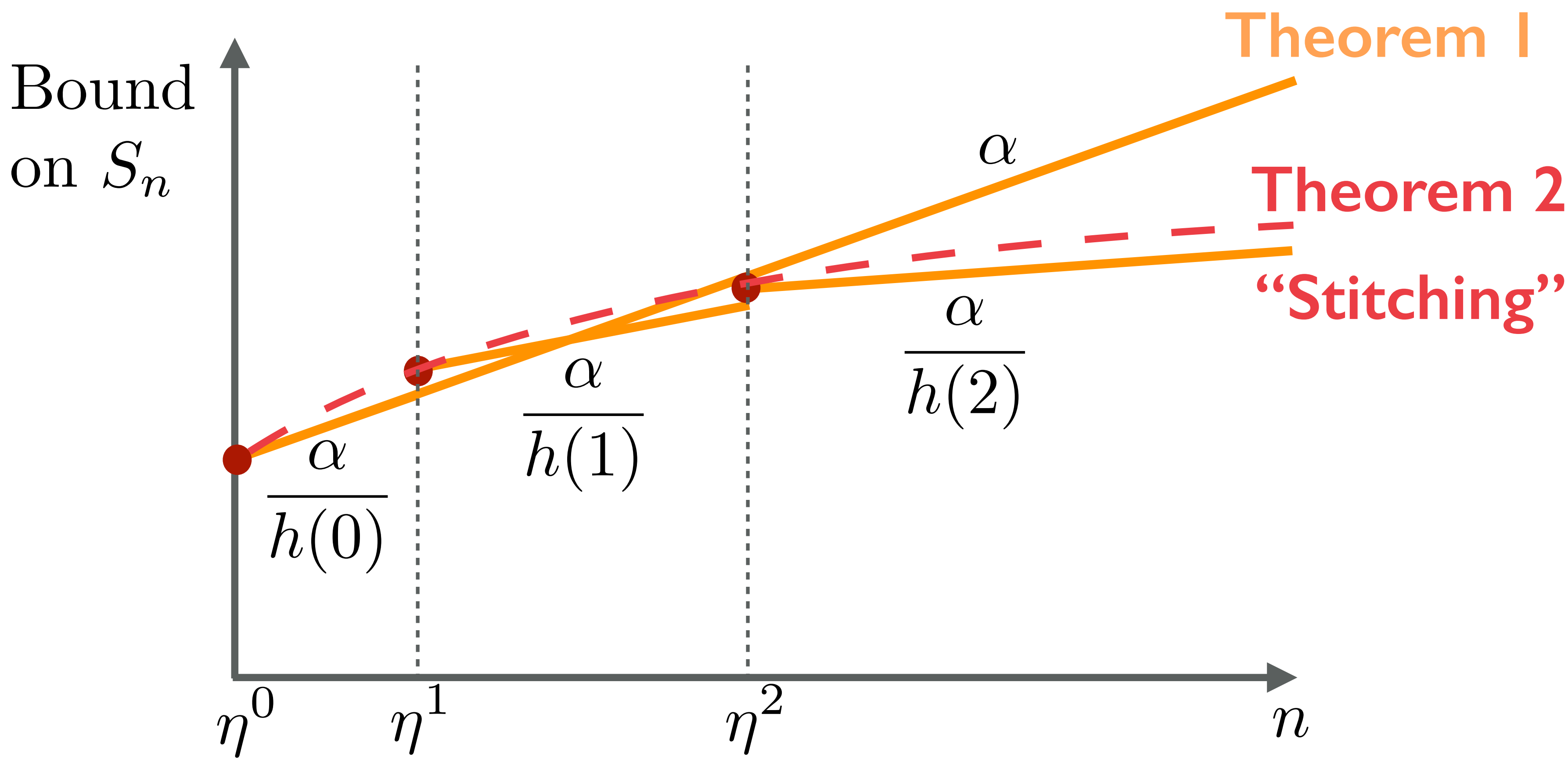
Mixture Method



Wald, Darling, Robbins, Siegmund, Lai, de la Pena, Howard et al.

Choose any constant $\eta > 1$ and any increasing function h such that $\sum_{k=0}^{\infty} 1/h(k) = 1$.

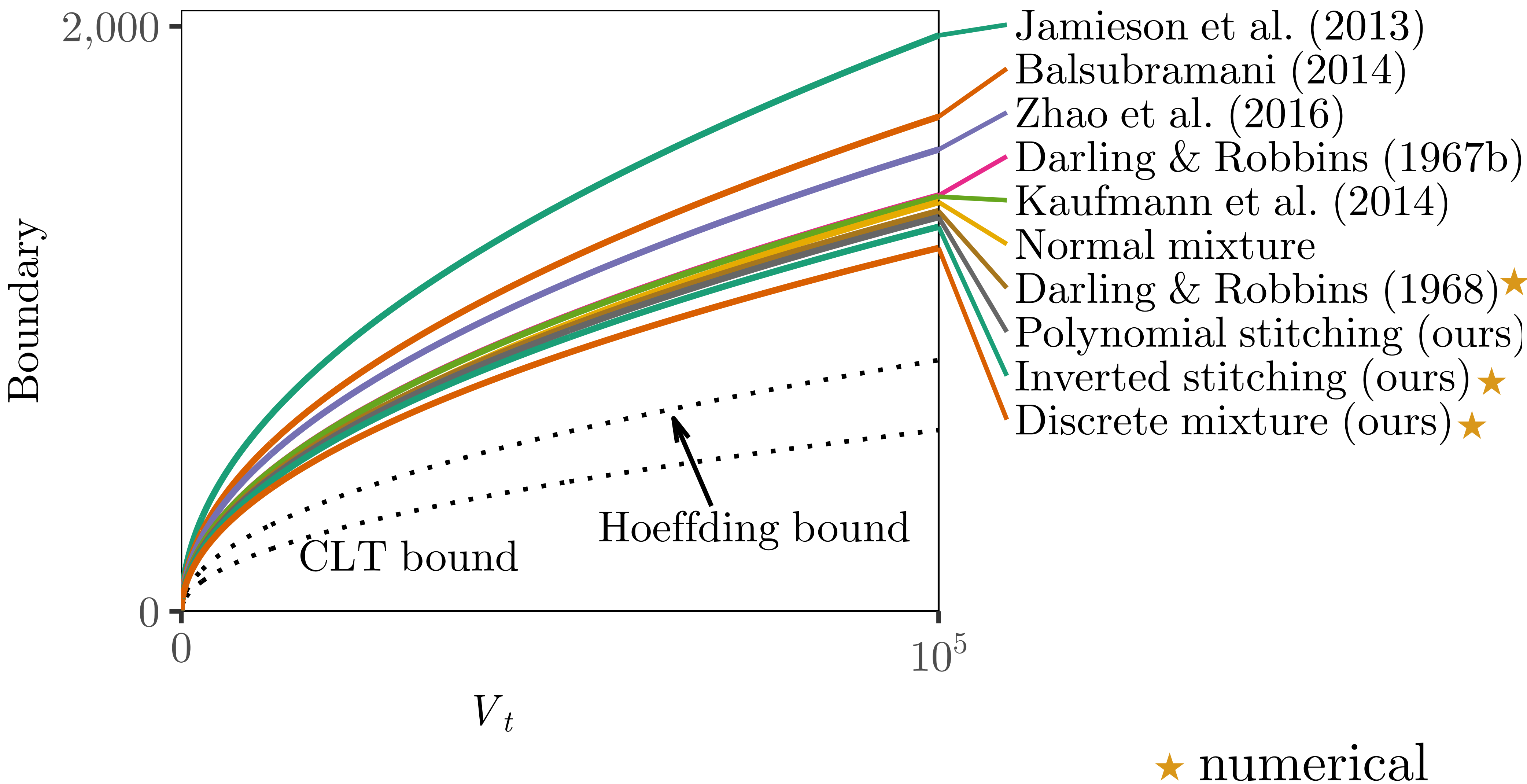
Eg: $h(k) \propto (k+1)^s$



Eg: X_i is 1-subGaussian 0-mean. Take $s = 1.4$, $\eta = 2$:

$$\mathbb{P}\left(\exists n \geq 1 : S_n \geq 1.71\sqrt{n\left(\log\log(2n) + 0.72\log\frac{5.19}{\alpha}\right)}\right) \leq \alpha.$$

A comparison for the (sub)Gaussian case



Application: testing if a coin is biased (or estimating its bias) by repeatedly tossing it

Choose any constant $\eta > 1$
and any increasing function h
such that $\sum_{k=0}^{\infty} 1/h(k) = 1$.

Then a confidence sequence for the bias is:

$$\frac{S_n}{n} \pm \frac{\eta^{1/4} + \eta^{-1/4}}{\sqrt{2}} \sqrt{\frac{\log h(\log_\eta n) + \log(2/\alpha)}{n}}.$$

Confidence sequence for fixed quantiles

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

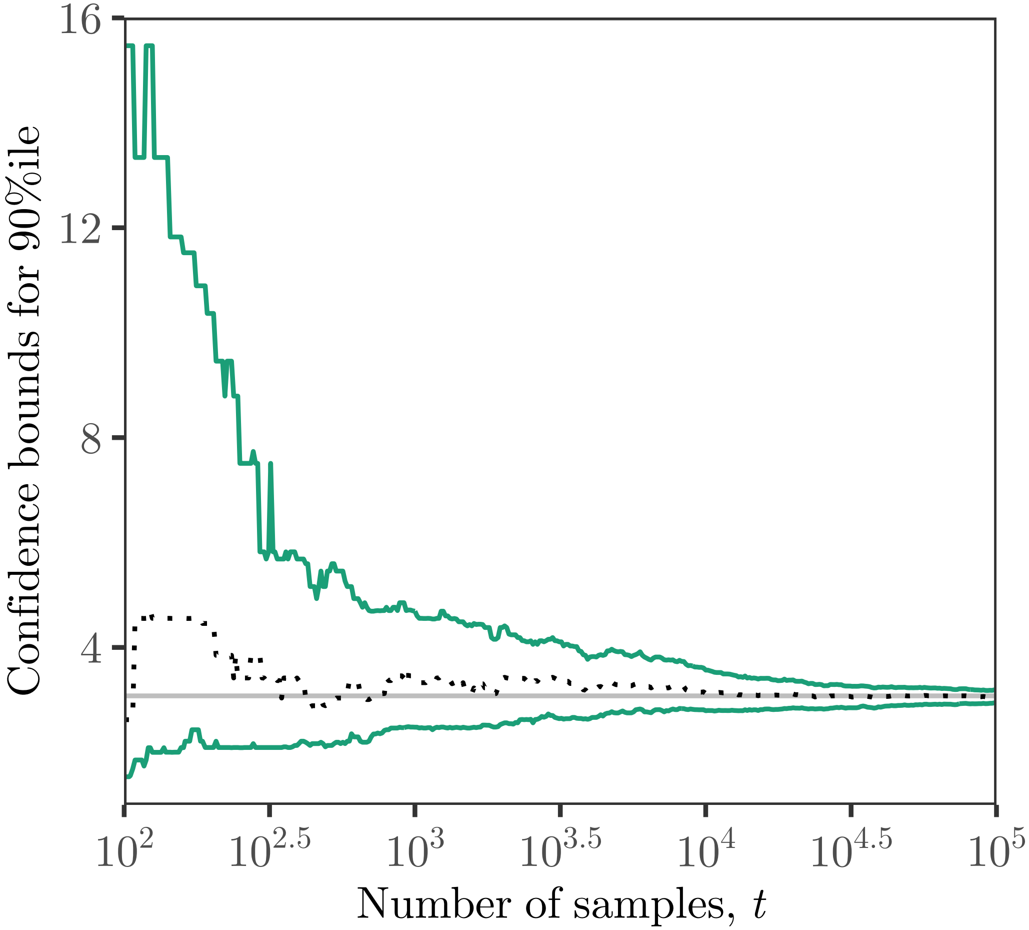
Then $\Pr(\forall t \in \mathbb{N} : \widehat{Q}_t(1/2 - u_t) \le Q(1/2) \le \widehat{Q}_t(1/2 + u_t)) \ge 1 - \alpha$.

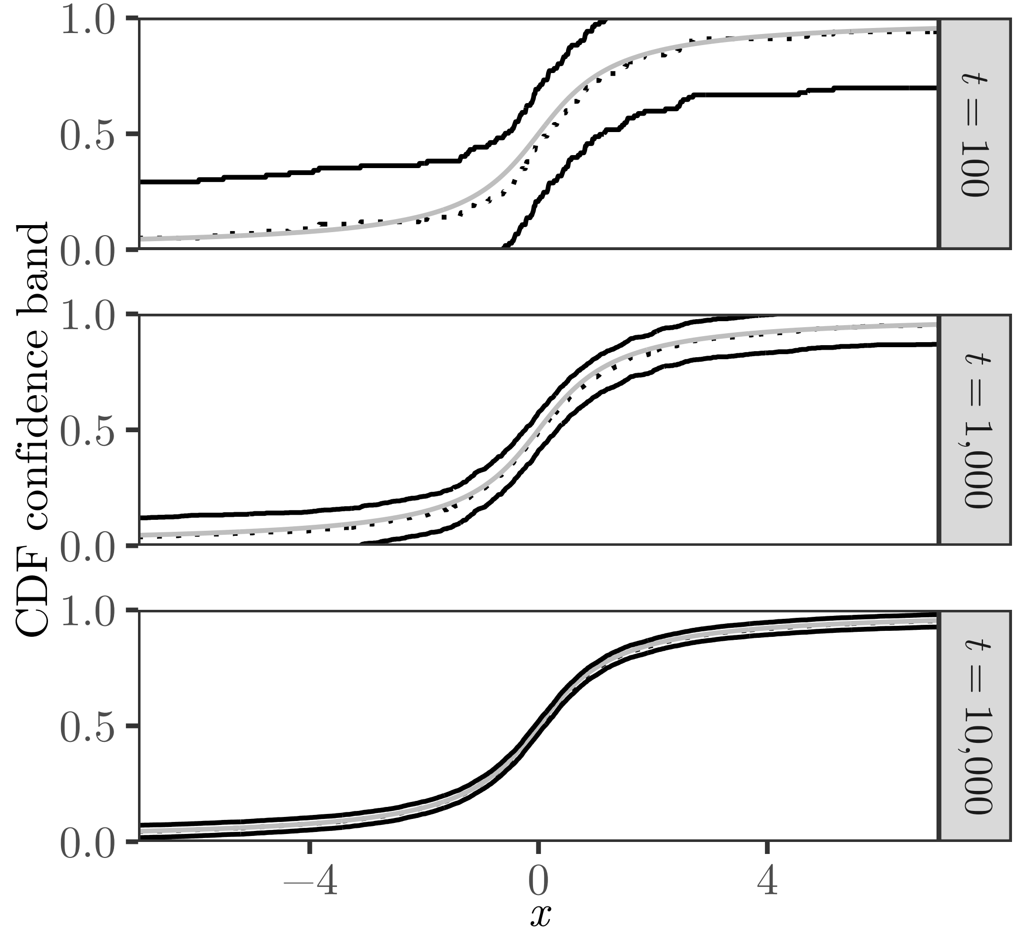
Confidence sequence for all quantiles simultaneously

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

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

Cauchy distribution





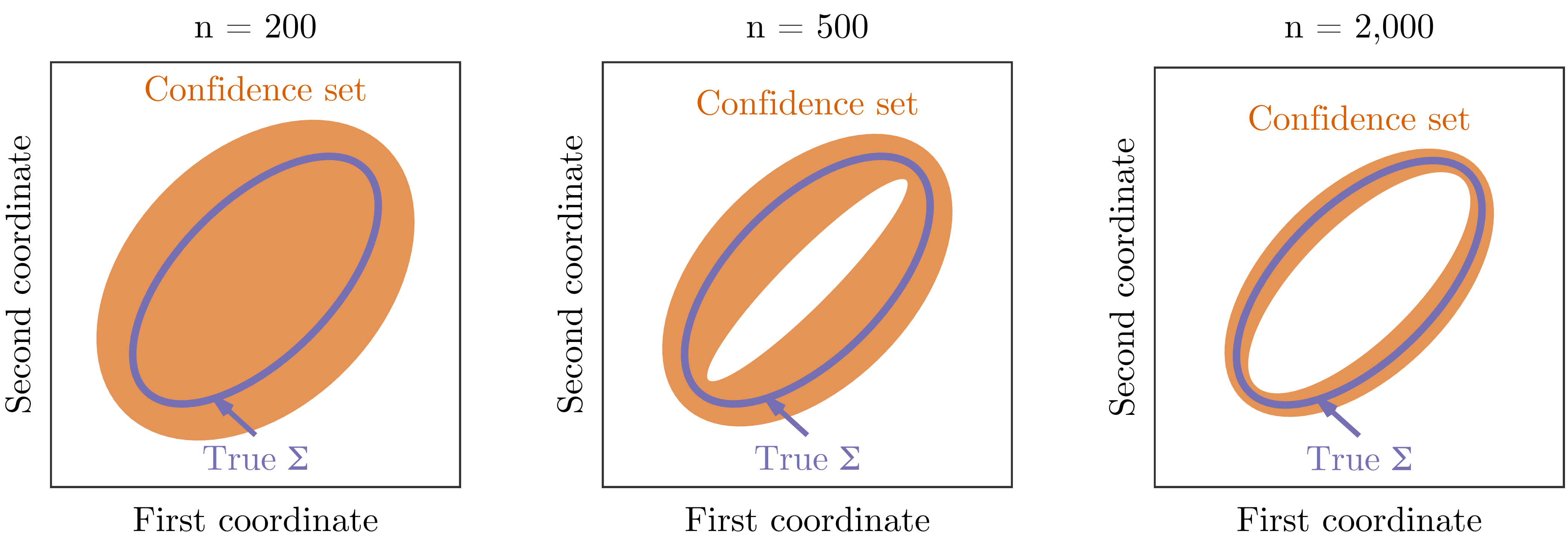
Only 0.9 quantile

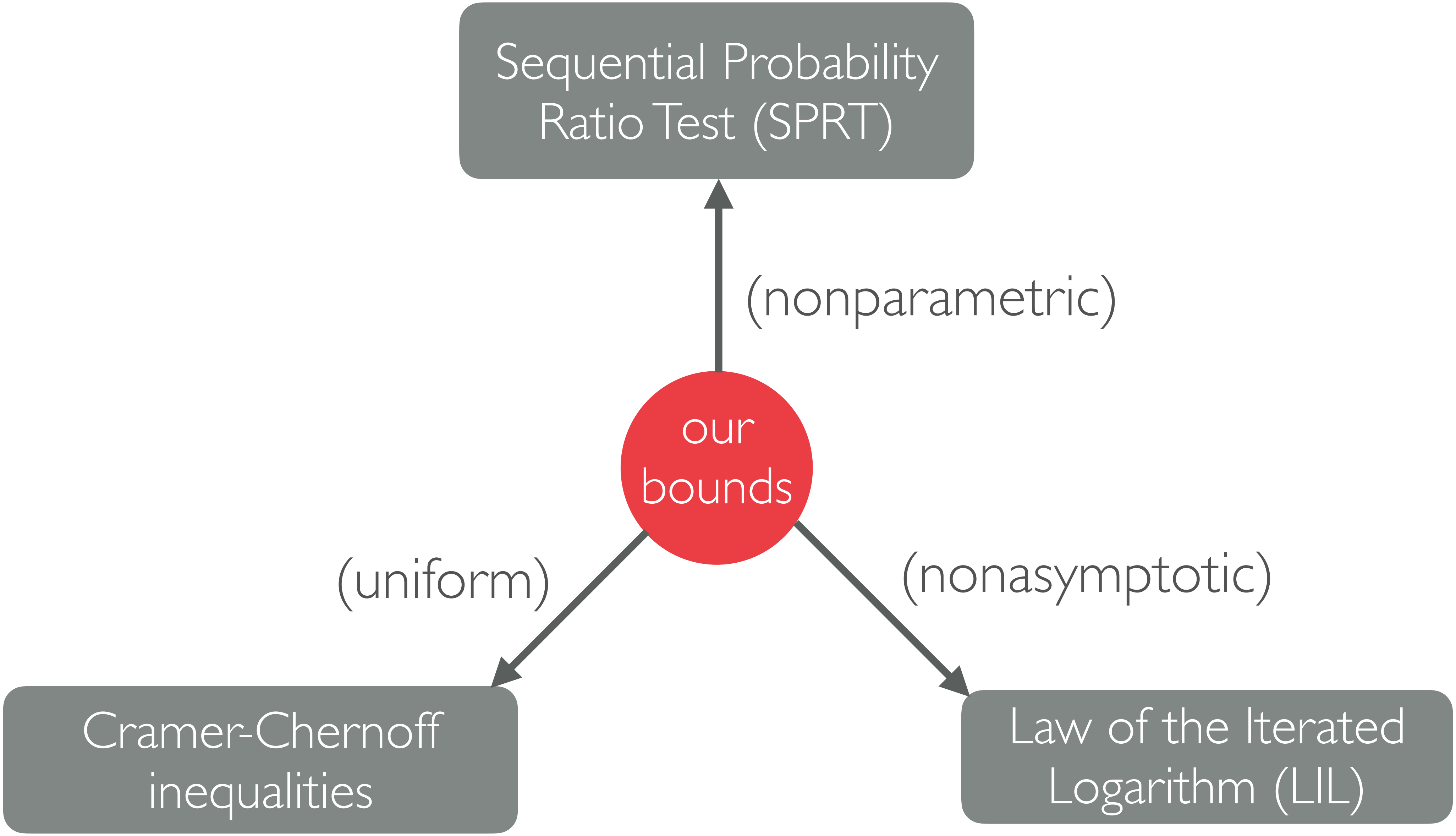
All quantiles simultaneously

Application: sequential covariance matrix estimation

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

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




Sequential Probability Ratio Test (SPRT)
(nonparametric)
our bounds
(uniform)
(nonasymptotic)
Cramer-Chernoff inequalities
Law of the Iterated Logarithm (LIL)

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.