Estimating means of bounded random variables by betting

Ian Waudby-Smith & Aaditya Ramdas
Carnegie Mellon University

Estimating means of bounded random variables by betting

Ian Waudby-Smith & Aaditya Ramdas
Carnegie Mellon University

JRSS-B 2023 (w/ discussion)

Given bounded rvs: $X_1, X_2, \ldots, X_n \in [0,1]$
with mean $\mathbb{E}(X_i) = \mu$,

Given bounded rvs: $X_1, X_2, \ldots, X_n \in [0,1]$
with mean $\mathbb{E}(X_i) = \mu$,

Goal: produce a set $C_n \equiv C(X_1, \ldots, X_n)$ s.t.

Given bounded rvs: $X_1, X_2, \ldots, X_n \in [0,1]$
with mean $\mathbb{E}(X_i) = \mu$,

Goal: produce a set $C_n \equiv C(X_1, \ldots, X_n)$ s.t.

$$\mathbb{P}(\mu \notin C_n) \leq \alpha$$

Given bounded rvs: $X_1, X_2, \ldots, X_n \in [0,1]$
with mean $\mathbb{E}(X_i) = \mu$,

Goal: produce a set $C_n \equiv C(X_1, \ldots, X_n)$ s.t.

$$\mathbb{P}(\mu \notin C_n) \leq \alpha \qquad \equiv \qquad \mathbb{P}(\mu \in C_n) \geq 1 - \alpha$$

Hoeffding's inequality (1963) provides one solution:

$$C_n^H := \frac{1}{n}\sum_{i=1}^{n} X_i \pm \sqrt{\frac{\log(2/\alpha)}{2n}}$$

Hoeffding's inequality (1963) provides one solution:

$$C_n^H := \frac{1}{n}\sum_{i=1}^{n} X_i \pm \sqrt{\frac{\log(2/\alpha)}{2n}}$$

No asymptotics, no parametric assumptions at all!

Hoeffding's inequality (1963) provides one solution:

$$C_n^H := \frac{1}{n}\sum_{i=1}^{n} X_i \pm \sqrt{\frac{\log(2/\alpha)}{2n}}$$

No asymptotics, no parametric assumptions at all!

The downside? Not very sharp,
especially for small variance $\sigma^2 := \mathrm{Var}(X_i)$.

Given bounded rvs: $X_1, X_2, \ldots, X_n \in [0,1]$ with mean $\mathbb{E}(X_i) = \mu$,

Goal: produce a set $C_n \equiv C(X_1, \ldots, X_n)$ s.t.

$$\mathbb{P}(\mu \notin C_n) \leq \alpha \quad \equiv \quad \mathbb{P}(\mu \in C_n) \geq 1 - \alpha,$$

Given bounded rvs: $X_1, X_2, \ldots, X_n \in [0,1]$
with mean $\mathbb{E}(X_i) = \mu$,

Goal: produce a set $C_n \equiv C(X_1, \ldots, X_n)$ s.t.

$$\mathbb{P}(\mu \notin C_n) \le \alpha \quad \equiv \quad \mathbb{P}(\mu \in C_n) \ge 1 - \alpha,$$

and **make C_n as sharp as possible.**

Our bound:

$$C_n := \left\{ m : \prod_{i=1}^{n} (1+\lambda_i(X_i - m)) < \frac{1}{\alpha} \right\}. \qquad \text{(design } \lambda_i \text{ later)}$$

Our bound:

$$C_n := \left\{ m : \prod_{i=1}^{n} (1+\lambda_i(X_i - m)) < \frac{1}{\alpha} \right\}. \qquad \text{(design } \lambda_i \text{ later)}$$

Hoeffding:

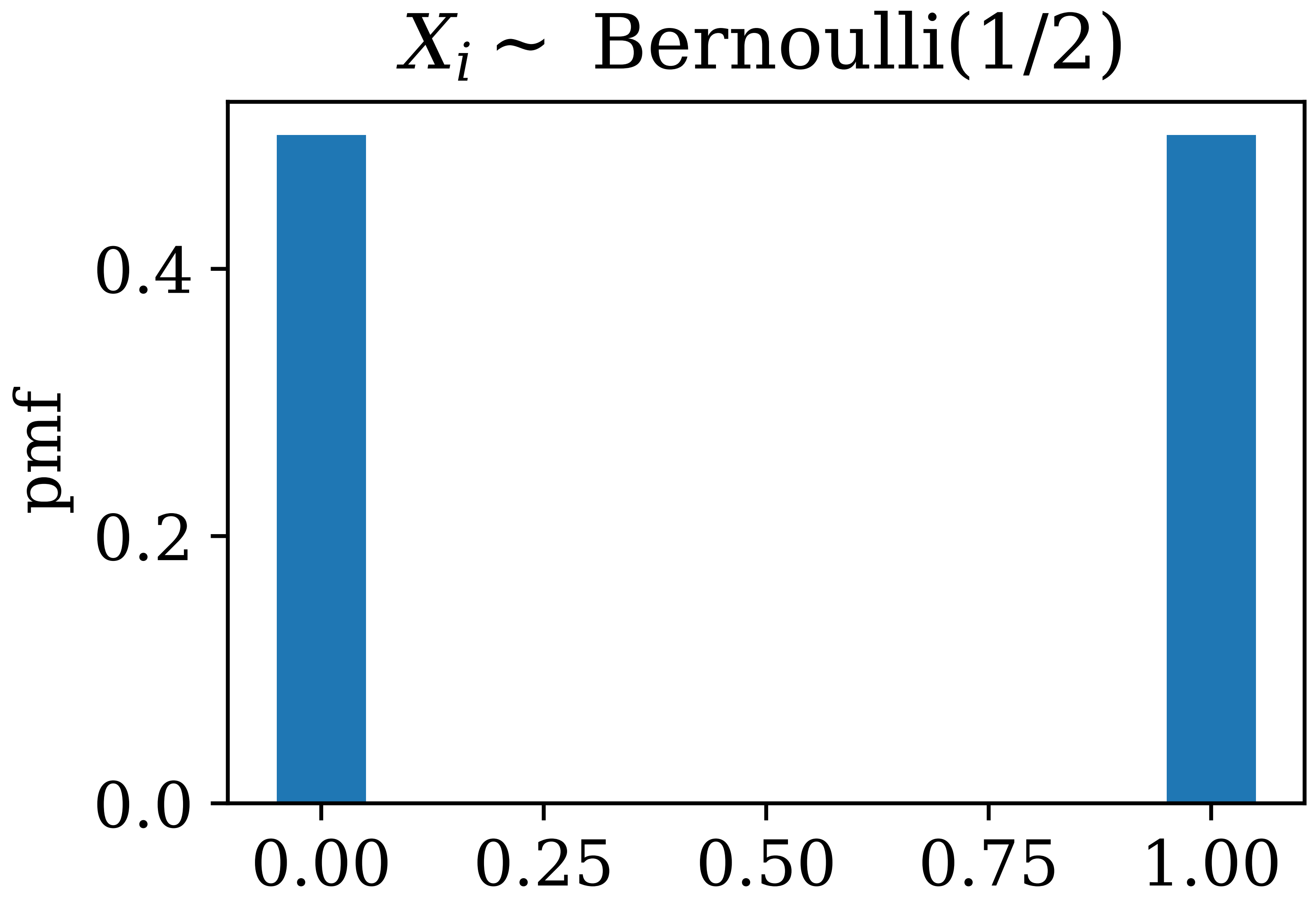
Our bound:

$$C_n := \left\{ m : \prod_{i=1}^{n} (1+\lambda_i(X_i - m)) < \frac{1}{\alpha} \right\}. \qquad \text{(design } \lambda_i \text{ later)}$$

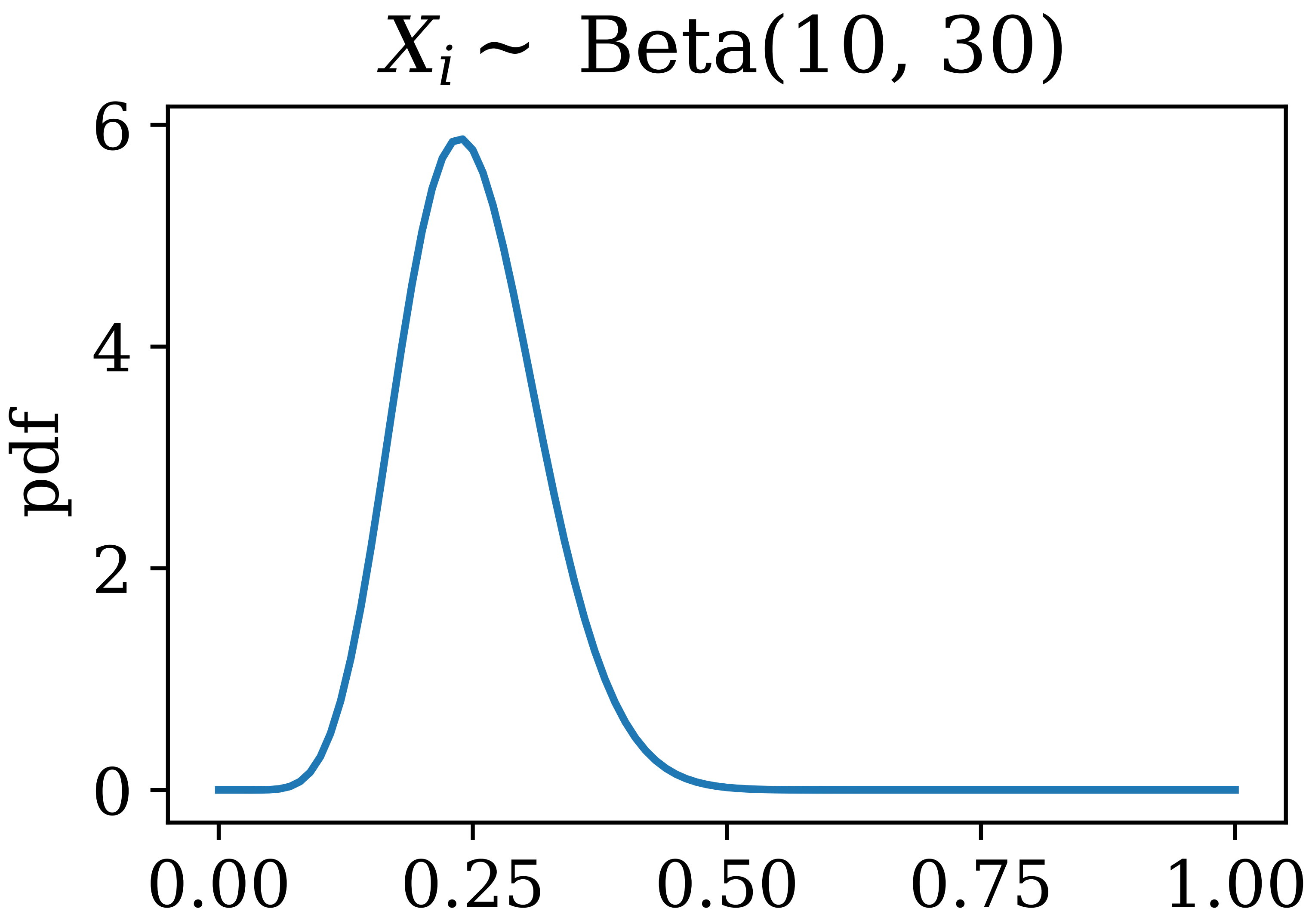
Hoeffding:

$$C_n^H := \left\{ m : \prod_{i=1}^{n} \exp\left\{\lambda(X_i - m) - \lambda^2/8\right\} < \frac{1}{\alpha} \right\}.$$

$\sigma^2 = 1/4$

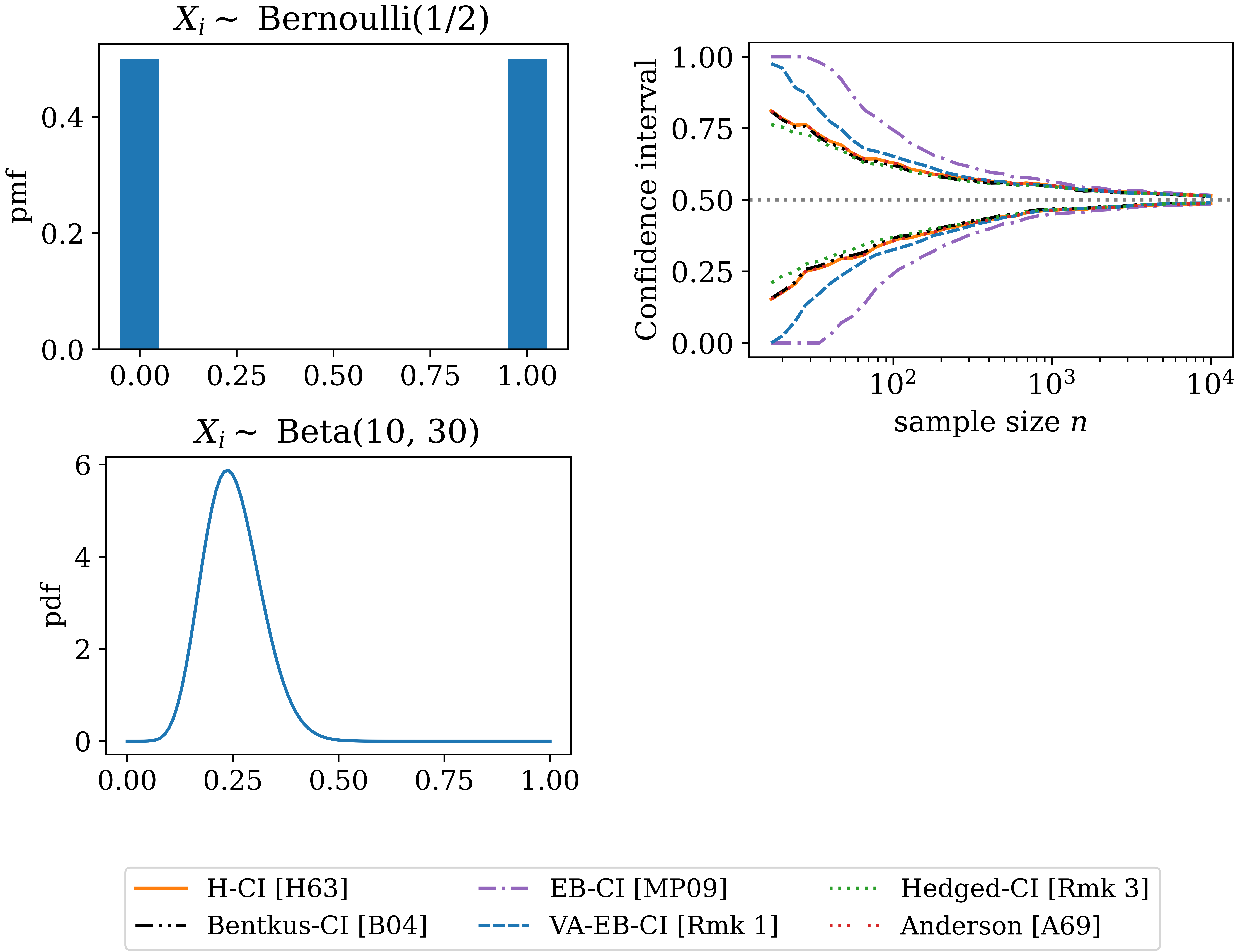


$\sigma^2 \approx 0.0046$

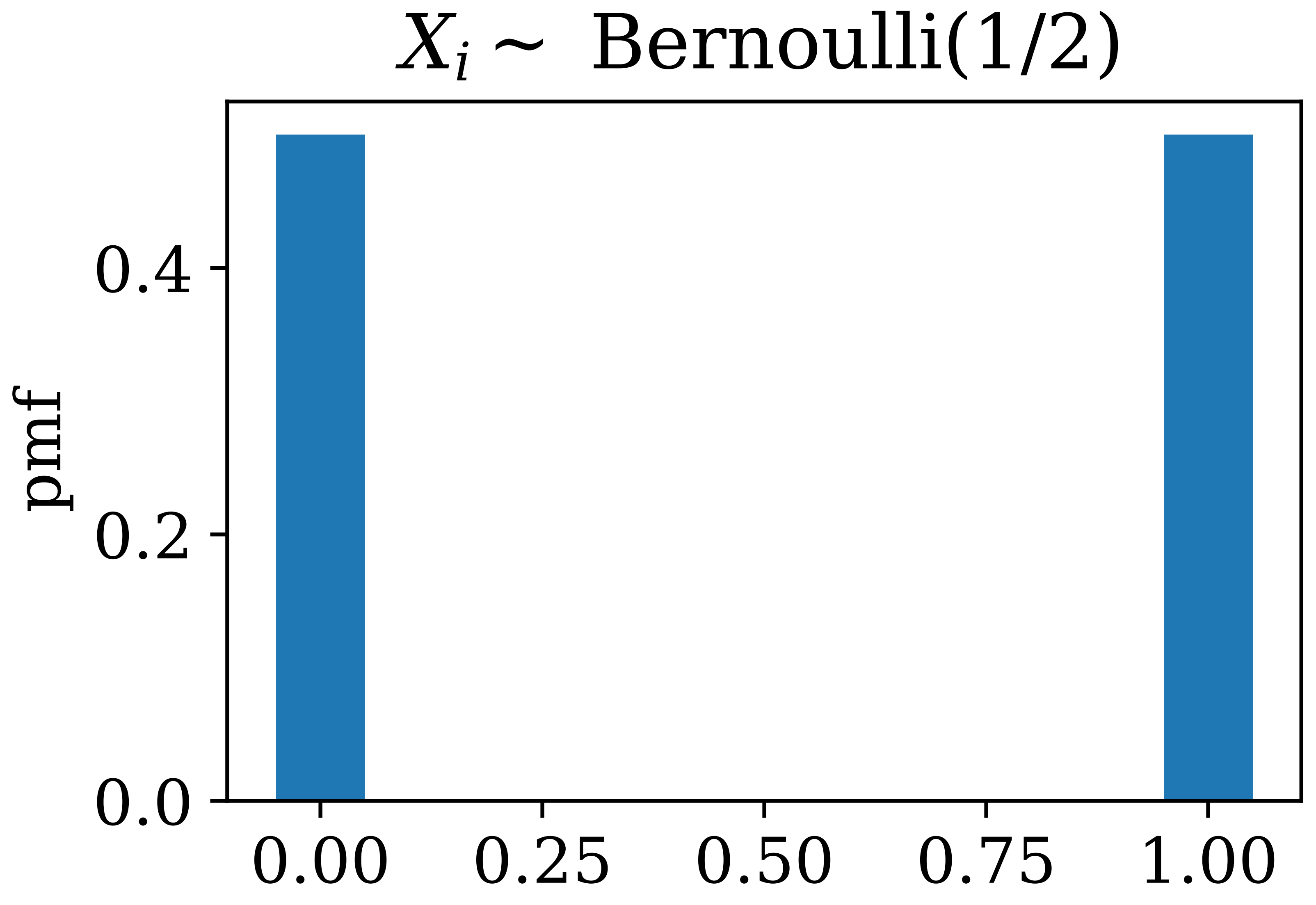


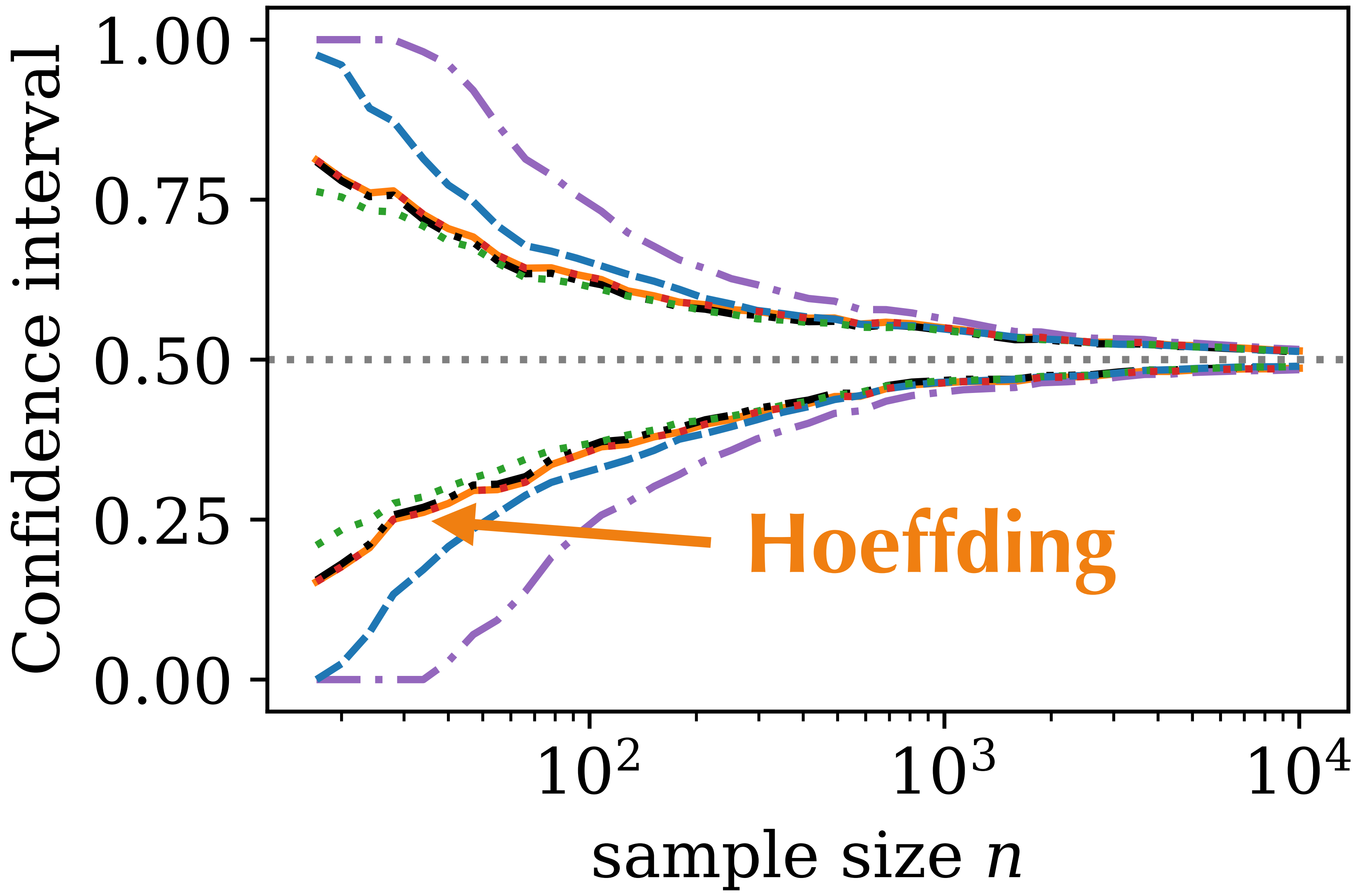
$\sigma^2 = 1/4$

$\sigma^2 \approx 0.0046$

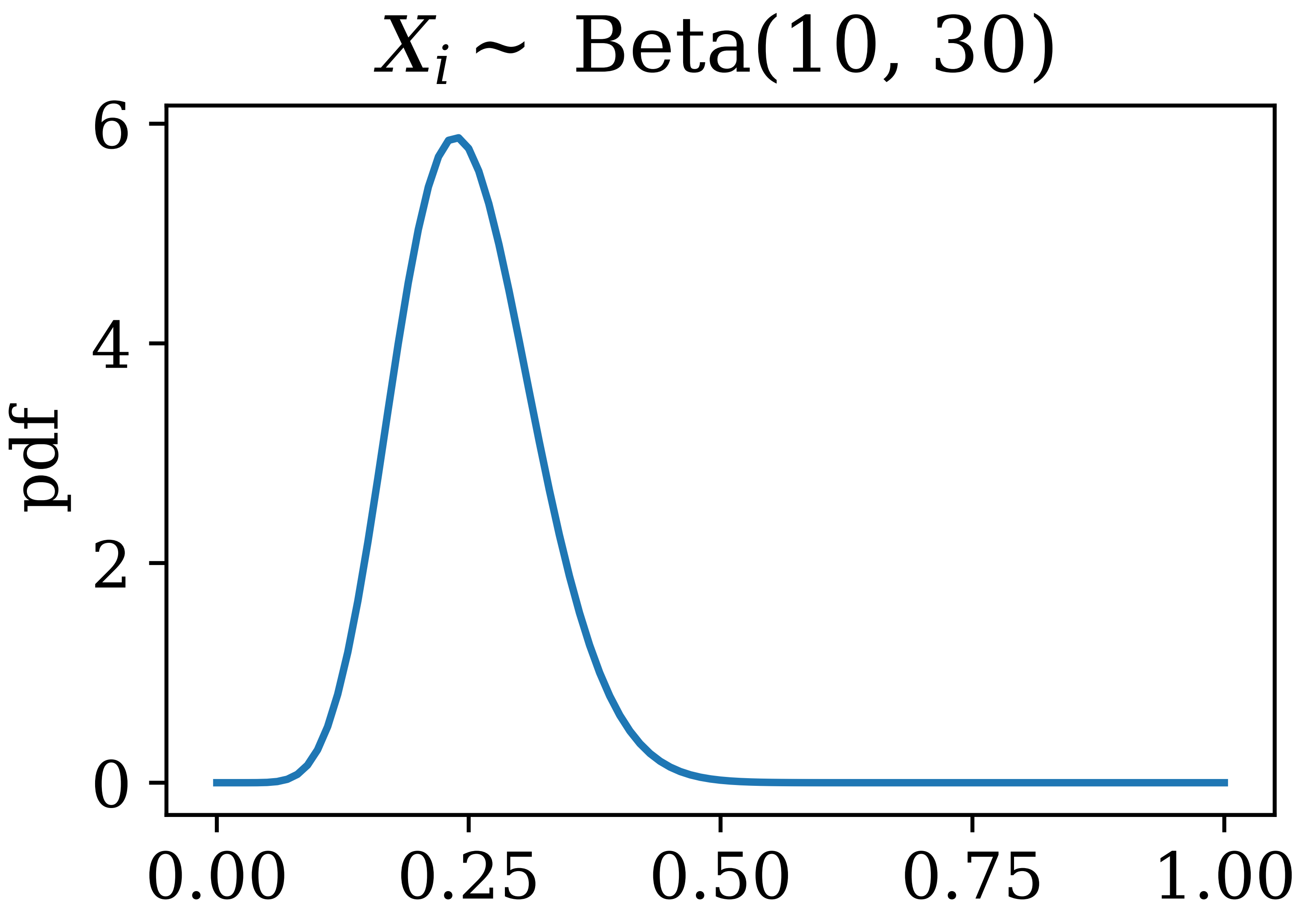


$\sigma^2 = 1/4$

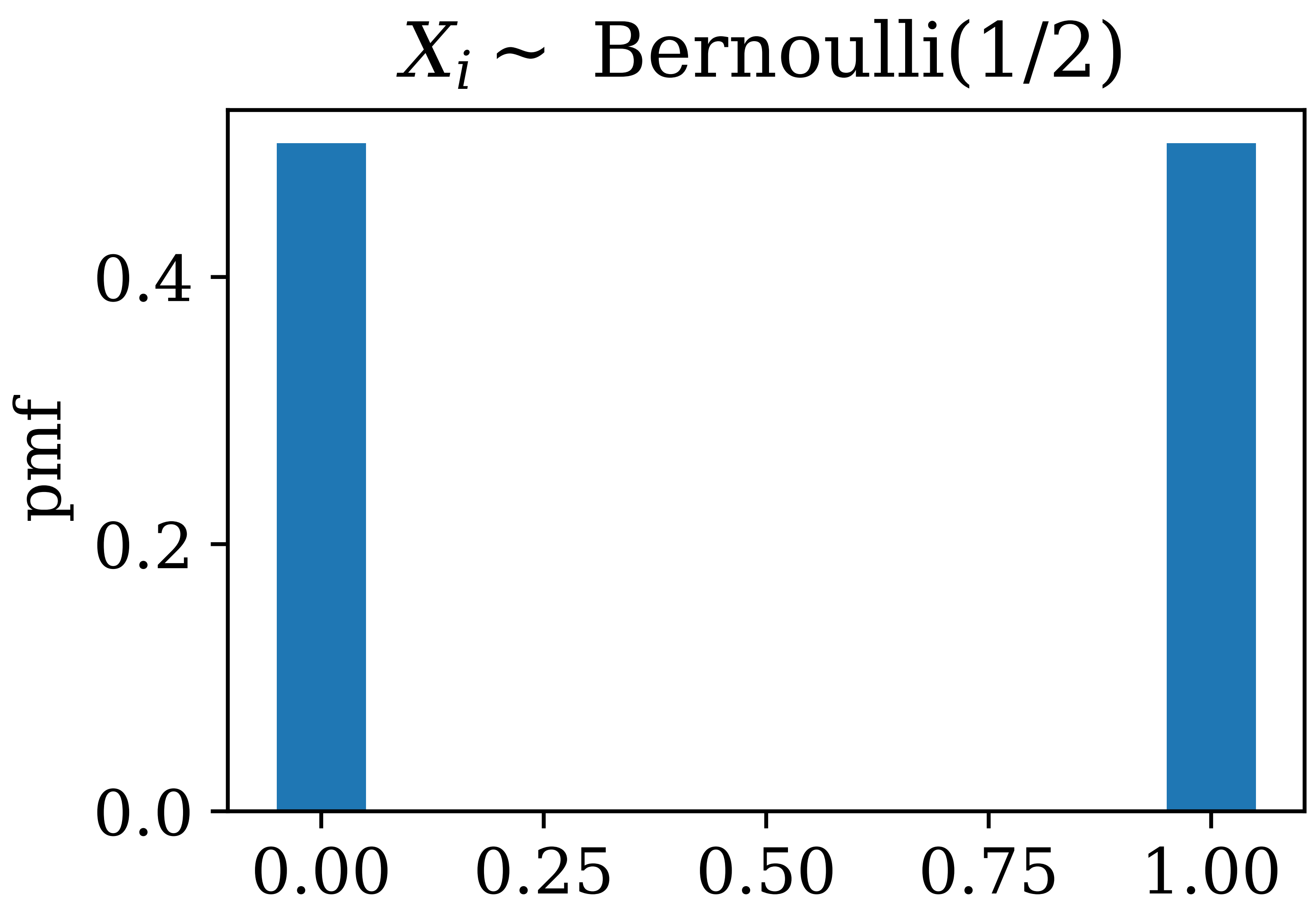


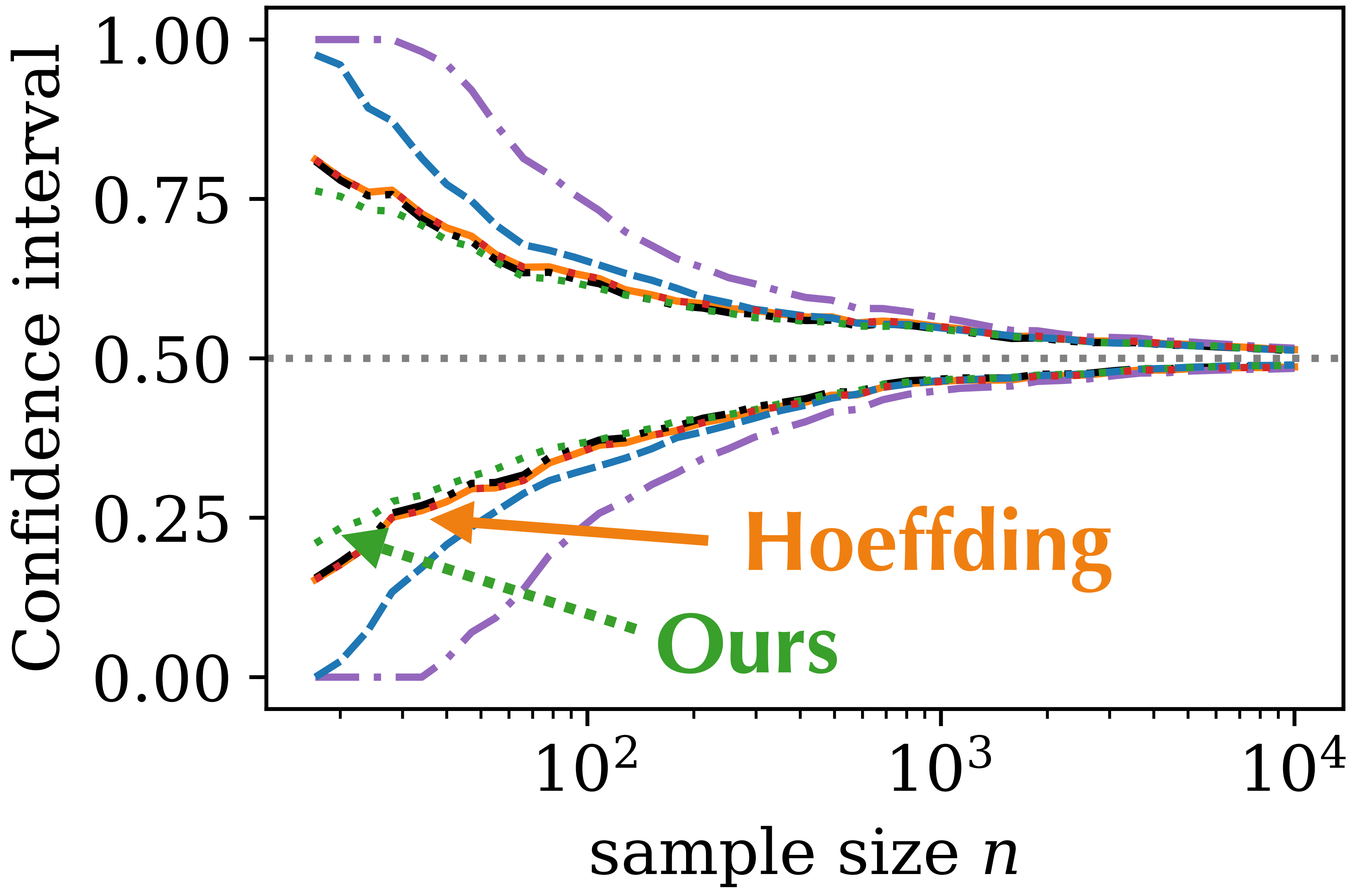


$\sigma^2 \approx 0.0046$

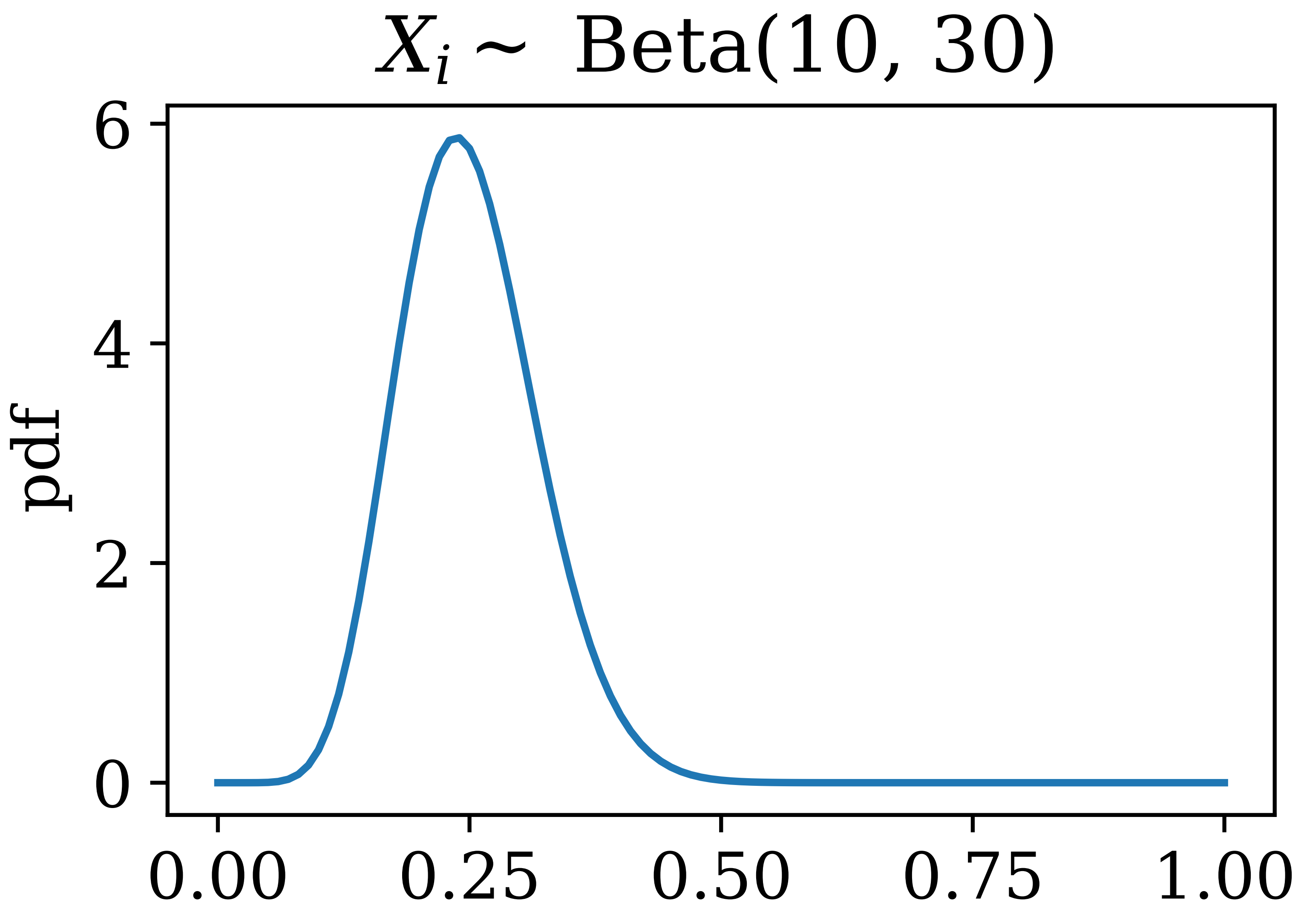


$\sigma^2 = 1/4$



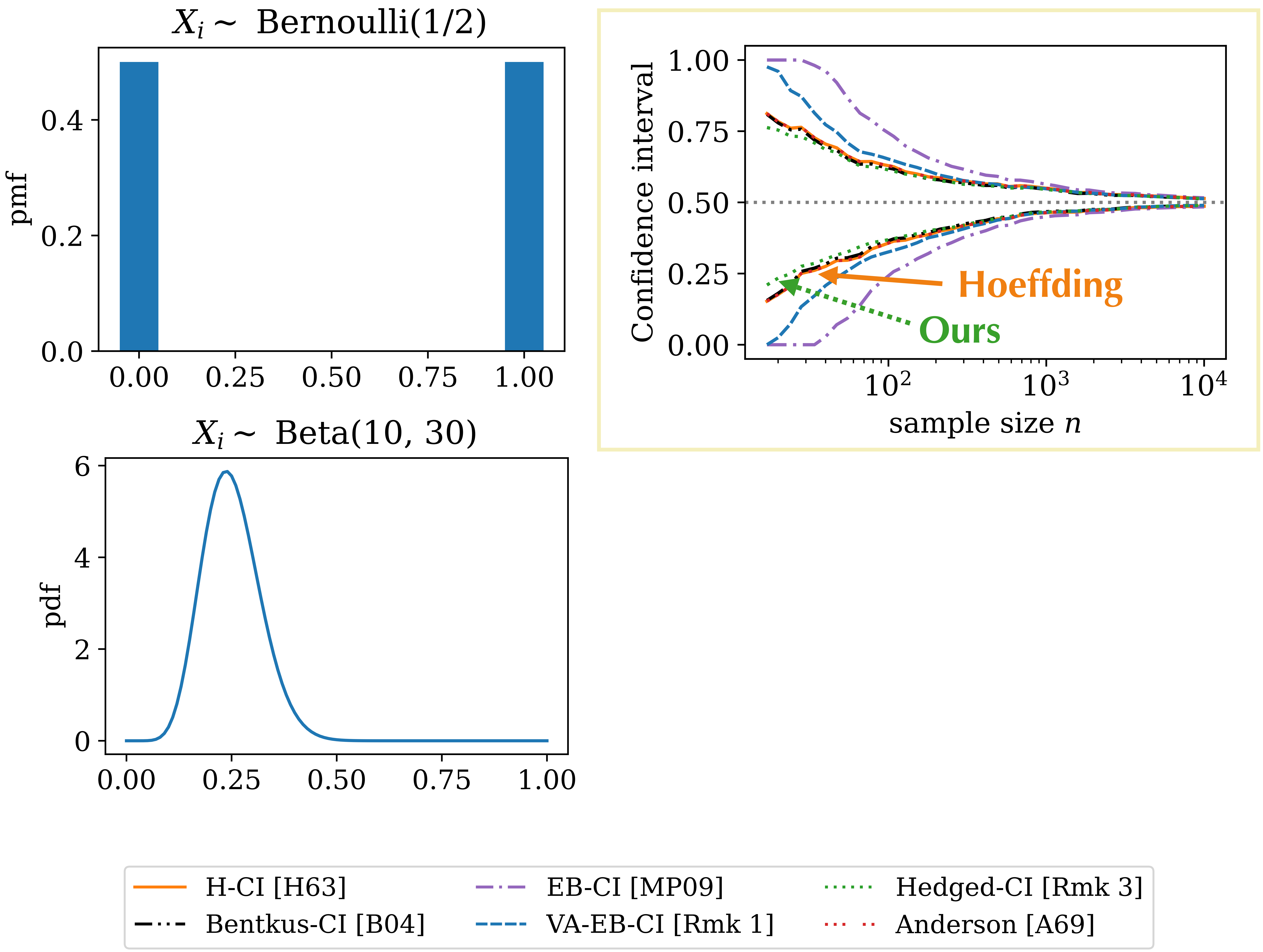


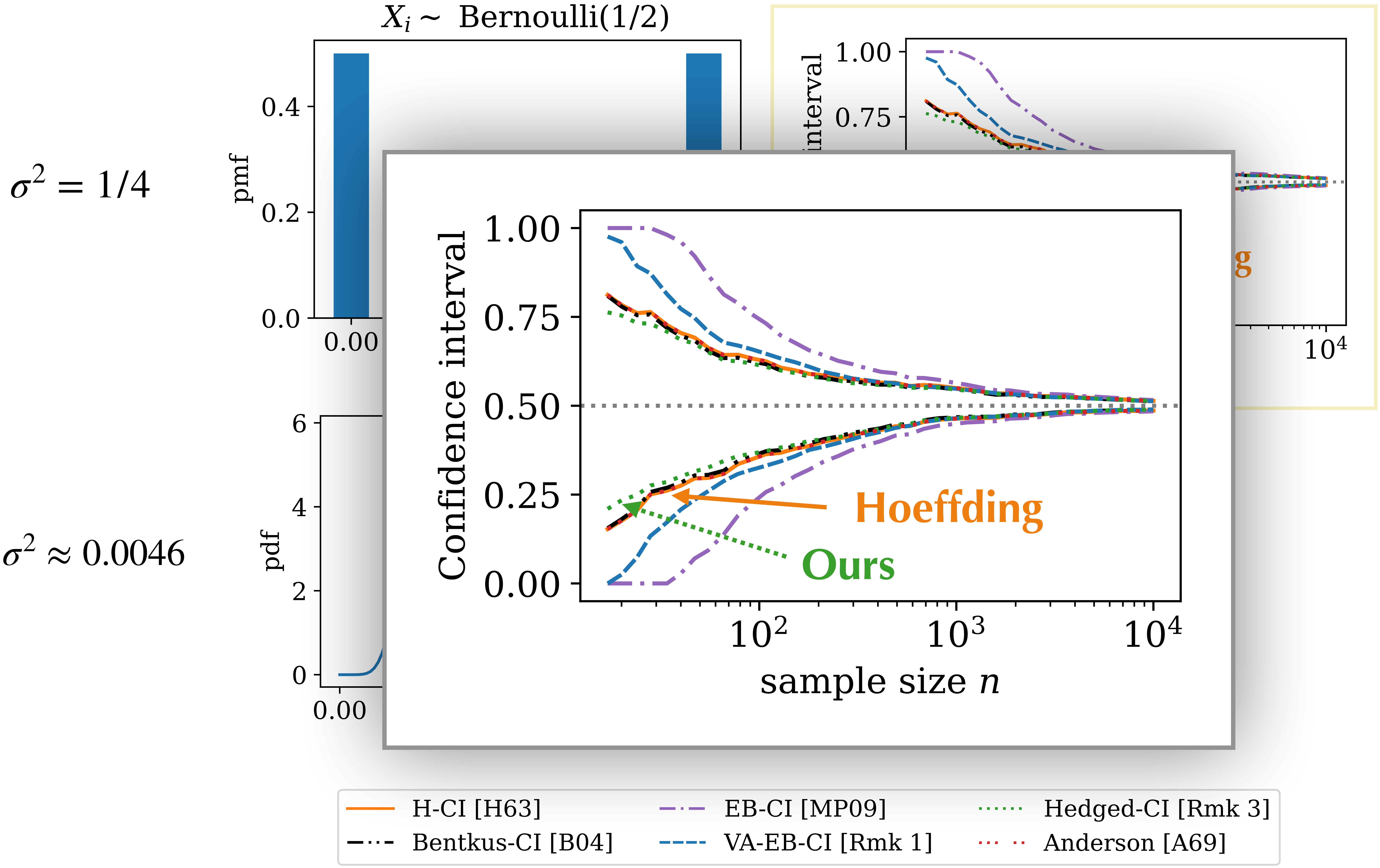
$\sigma^2 \approx 0.0046$



$\sigma^2 = 1/4$

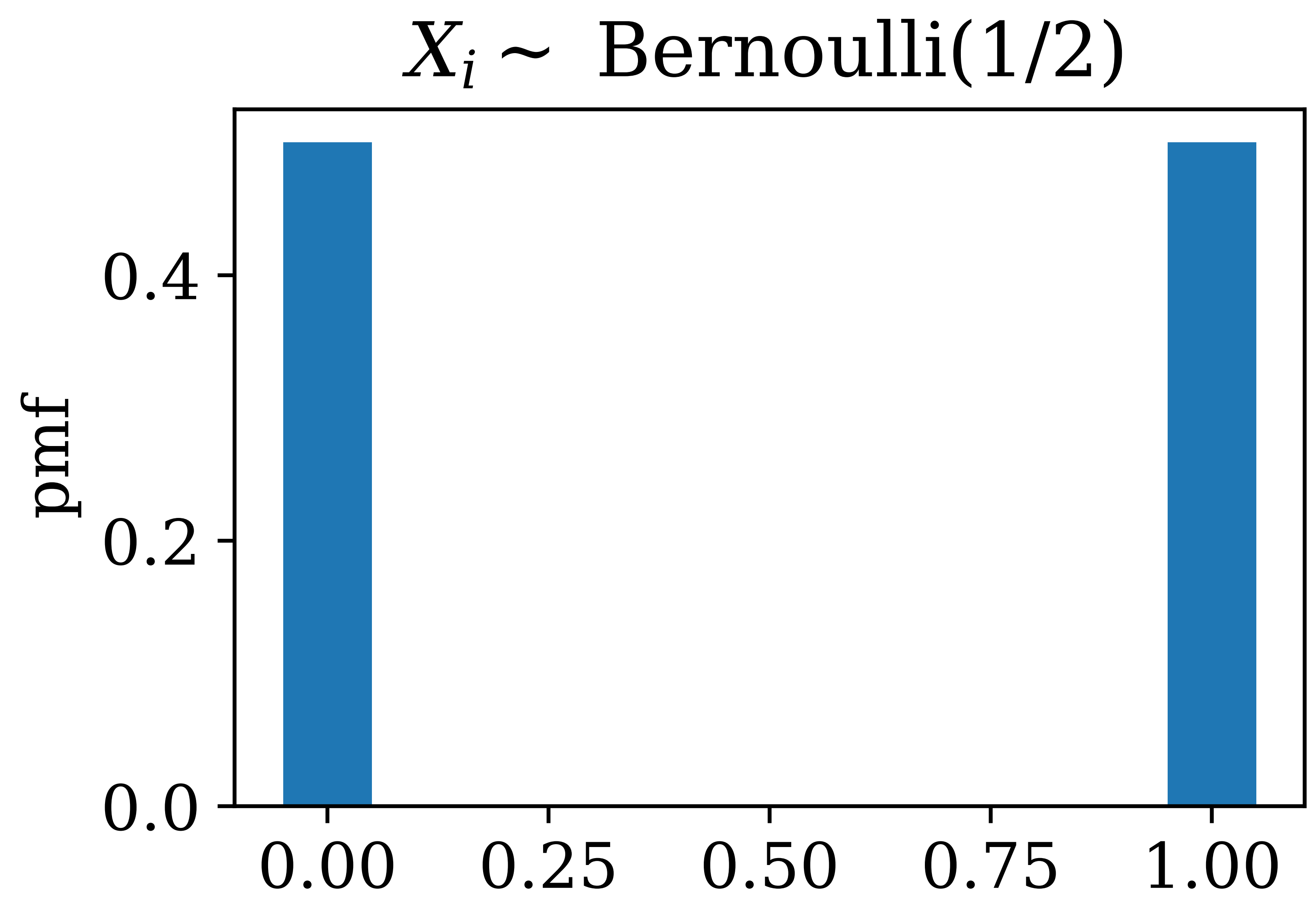
$\sigma^2 \approx 0.0046$

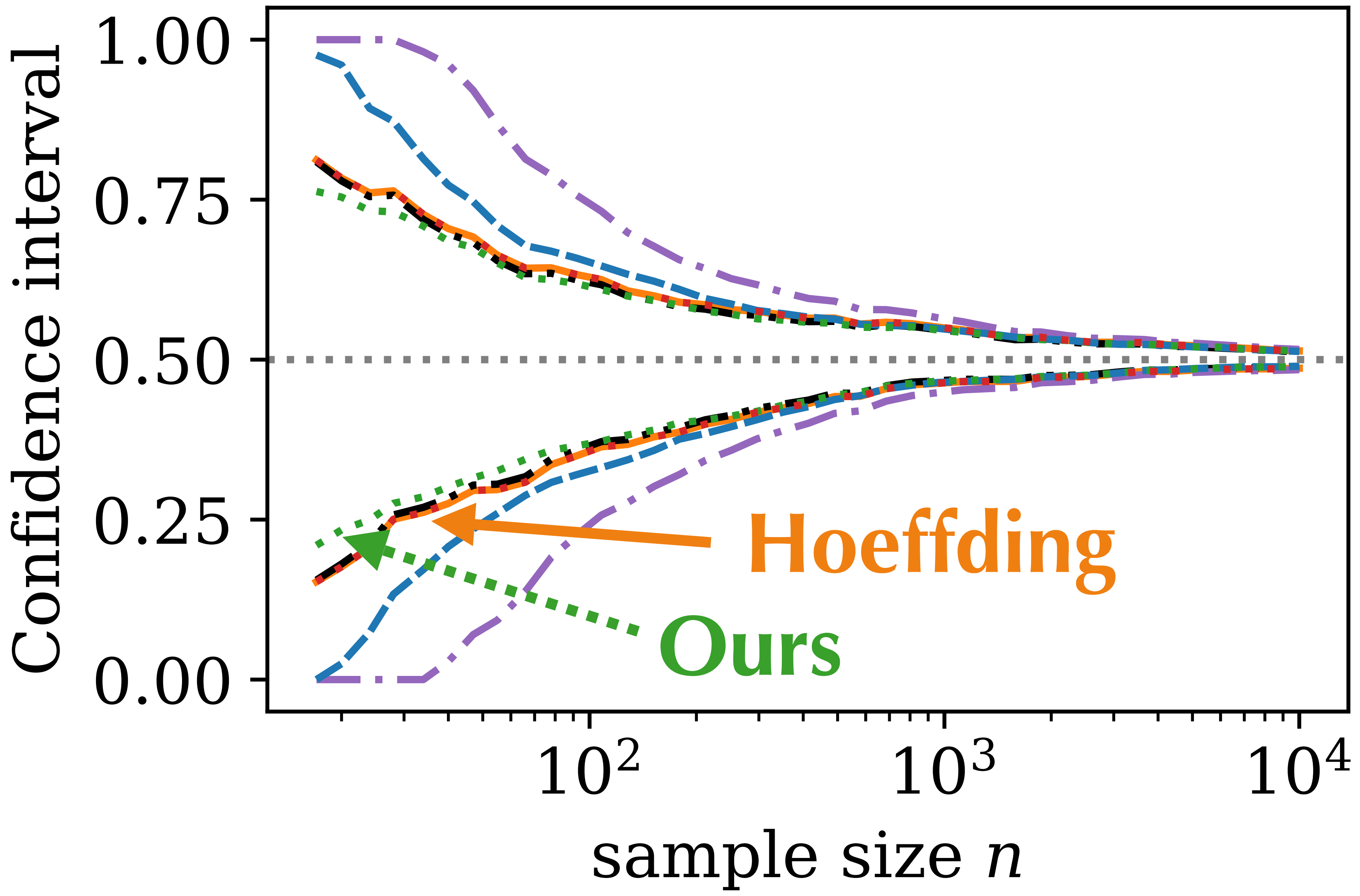




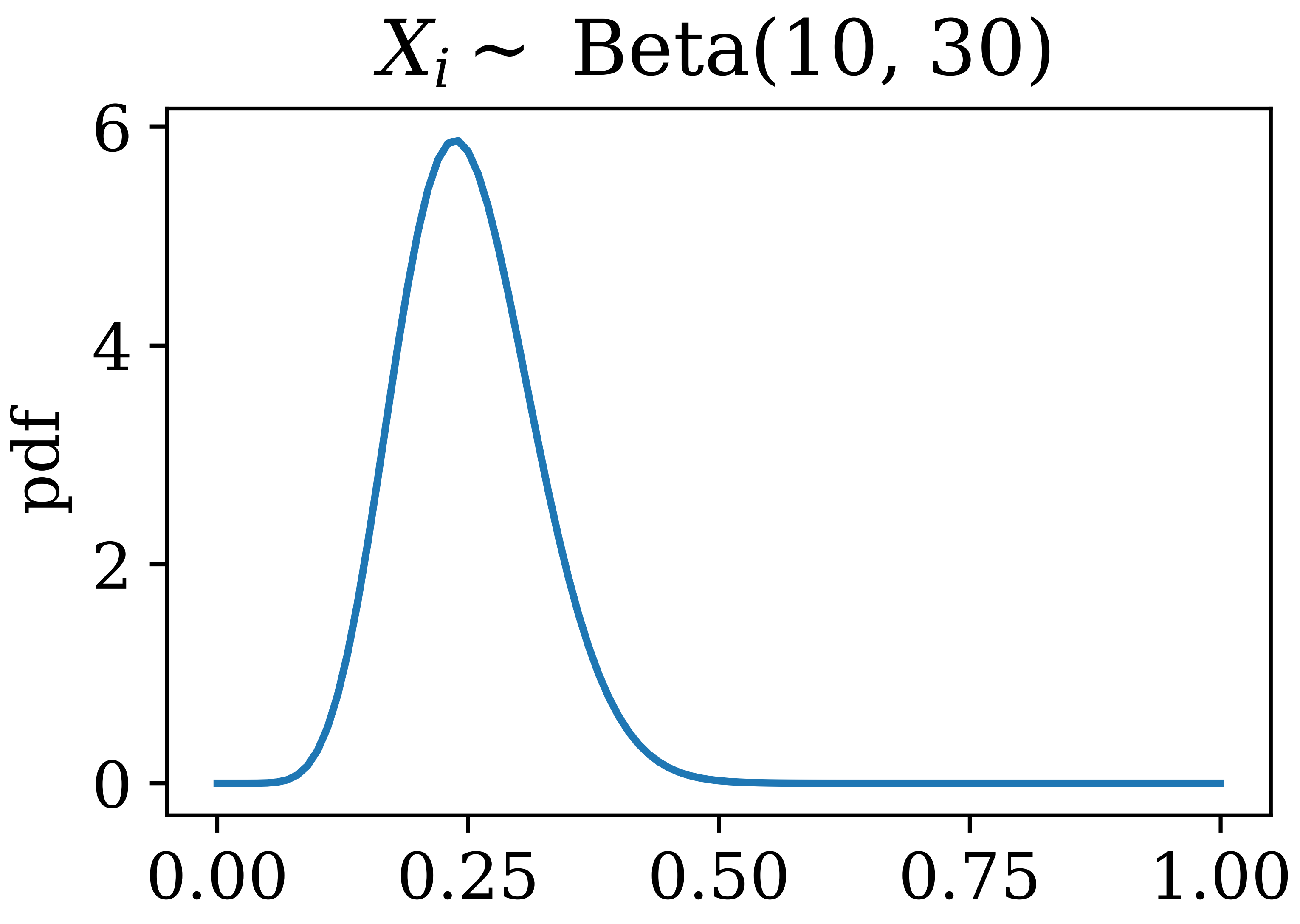
$\sigma^2 = 1/4$
$X_i \sim$ Bernoulli(1/2)
pmf
$\sigma^2 \approx 0.0046$
pdf
Confidence interval
sample size n
Hoeffding
Ours
H-CI [H63]
Bentkus-CI [B04]
EB-CI [MP09]
VA-EB-CI [Rmk 1]
Hedged-CI [Rmk 3]
Anderson [A69]

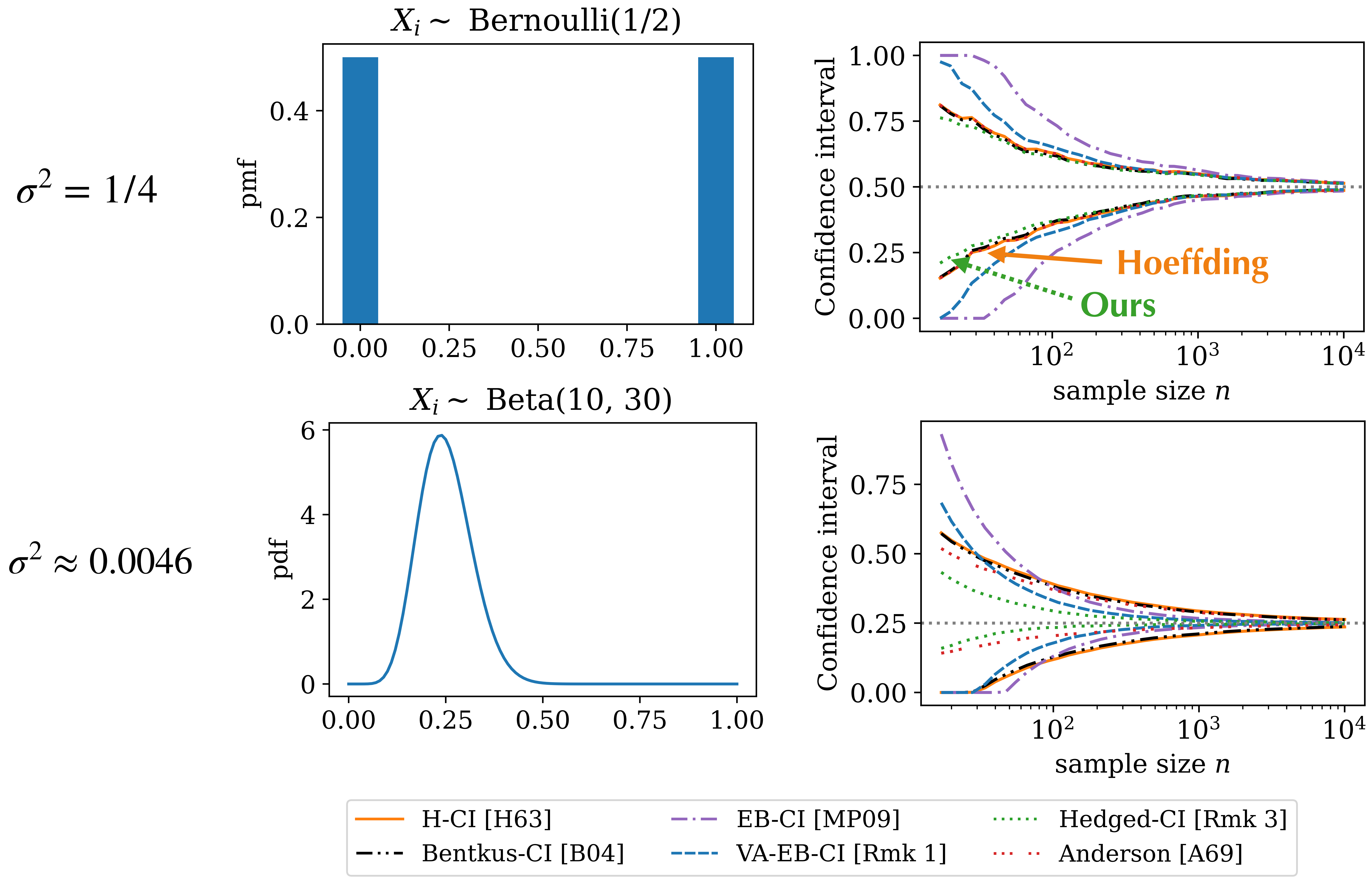
$\sigma^2 = 1/4$

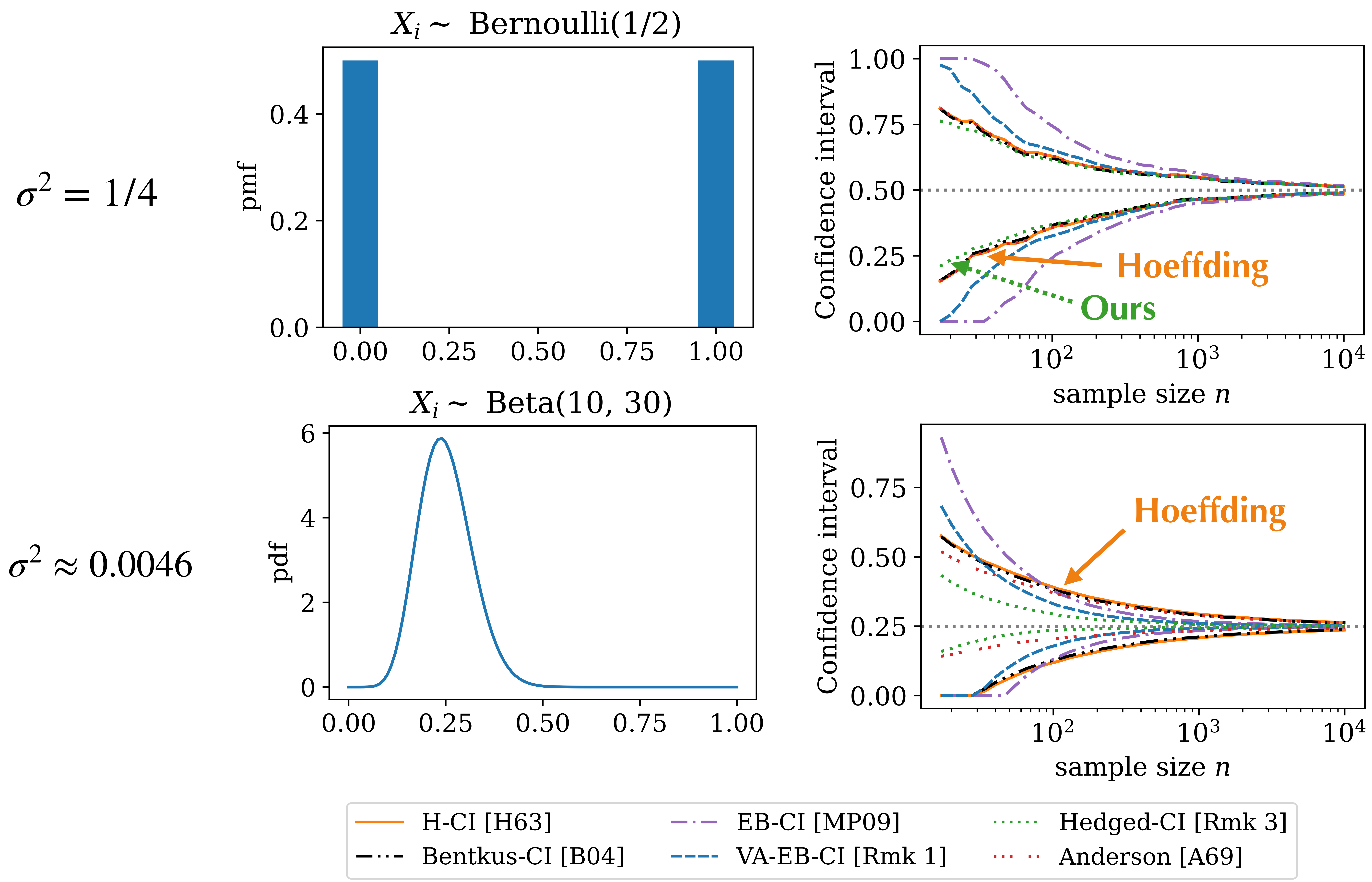




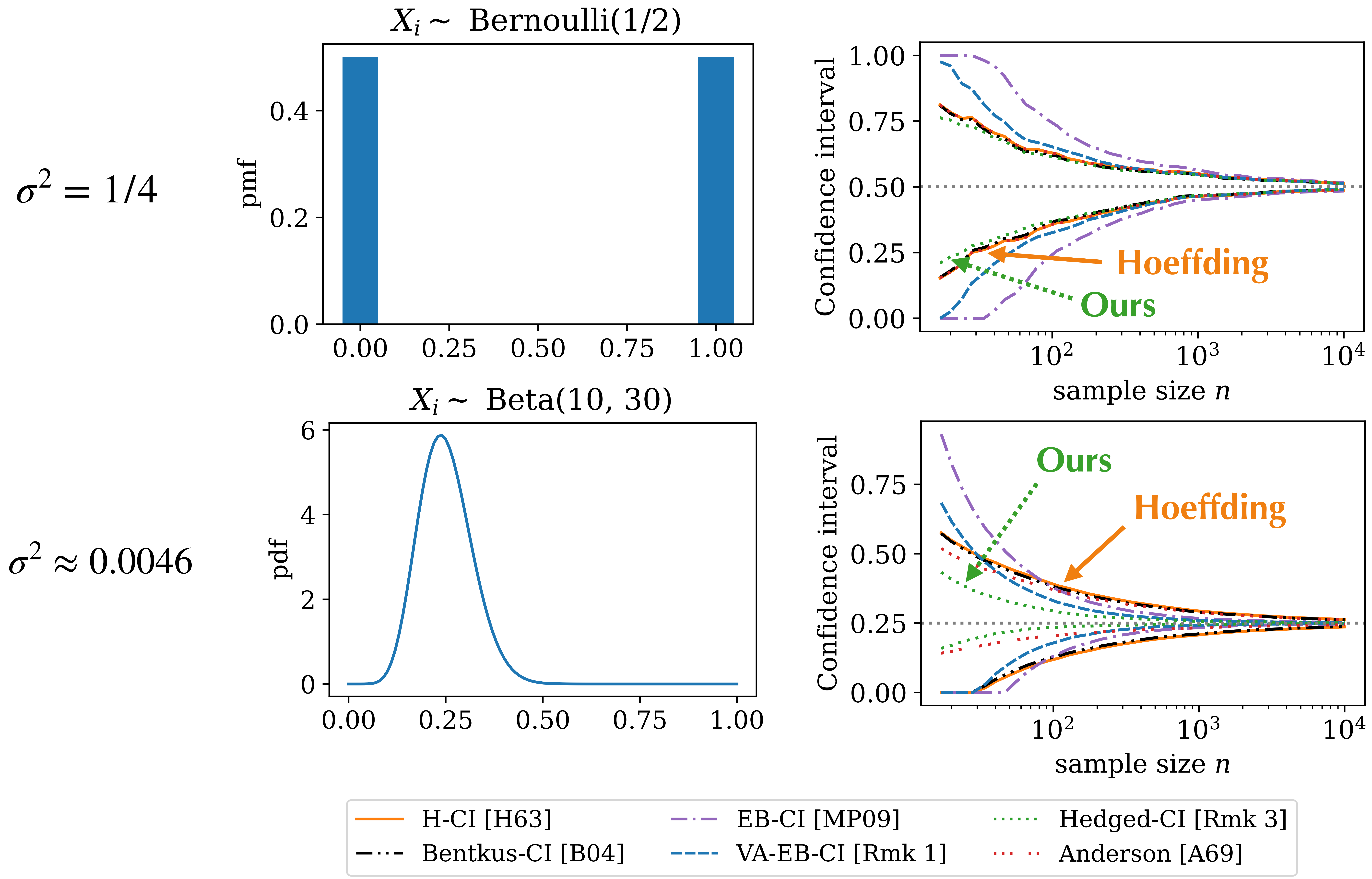
$\sigma^2 \approx 0.0046$

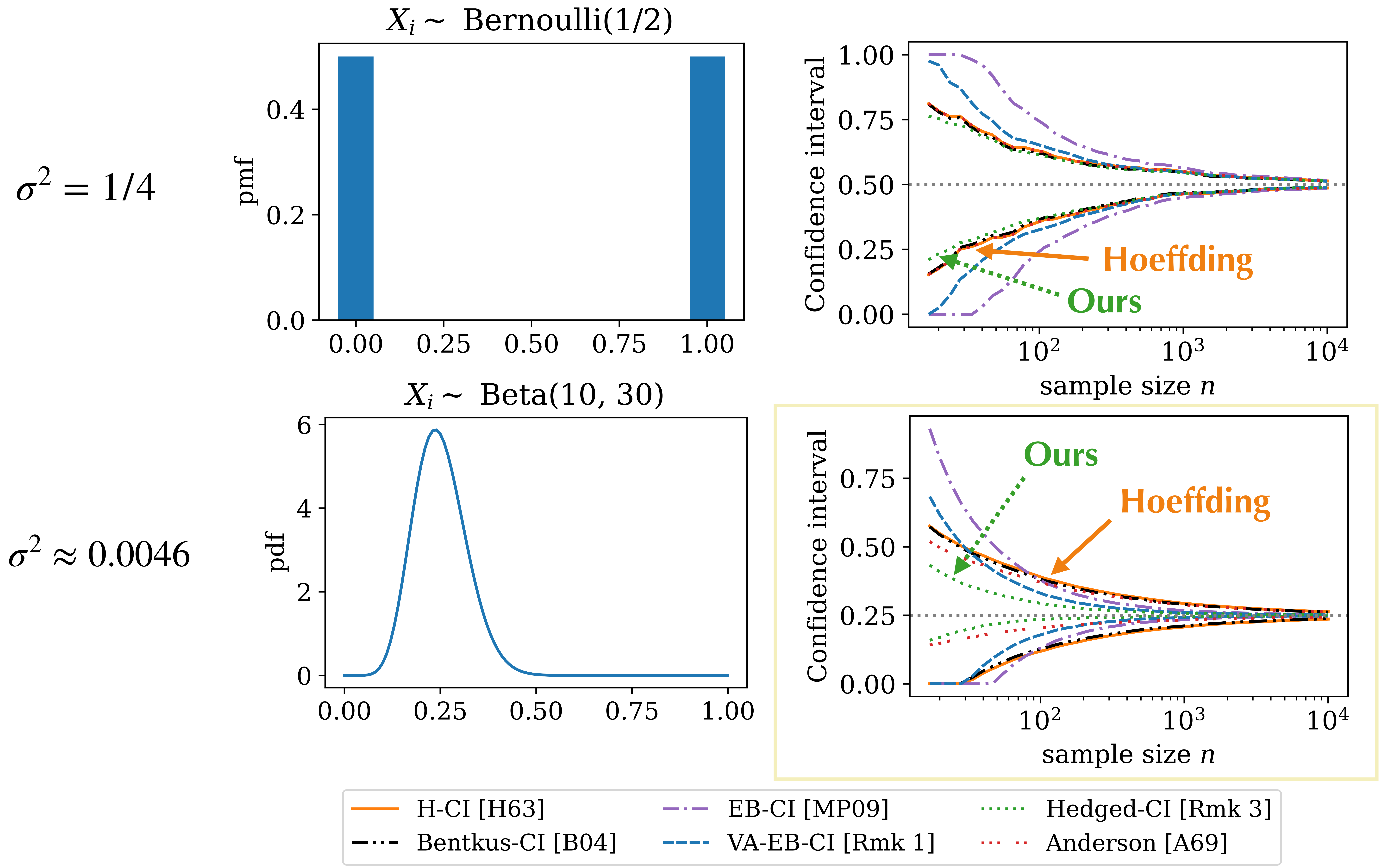



$\sigma^2 = 1/4$
$X_i \sim$ Bernoulli(1/2)
pmf
0.0
0.2
0.4
0.00
0.25
0.50
0.75
1.00
Confidence interval
0.00
0.25
0.50
0.75
1.00
Hoeffding
Ours
10^2
10^3
10^4
sample size n
$\sigma^2 \approx 0.0046$
$X_i \sim$ Beta(10, 30)
pdf
0
2
4
6
0.00
0.25
0.50
0.75
1.00
Confidence interval
0.00
0.25
0.50
0.75
10^2
10^3
10^4
sample size n
H-CI [H63]
Bentkus-CI [B04]
EB-CI [MP09]
VA-EB-CI [Rmk 1]
Hedged-CI [Rmk 3]
Anderson [A69]

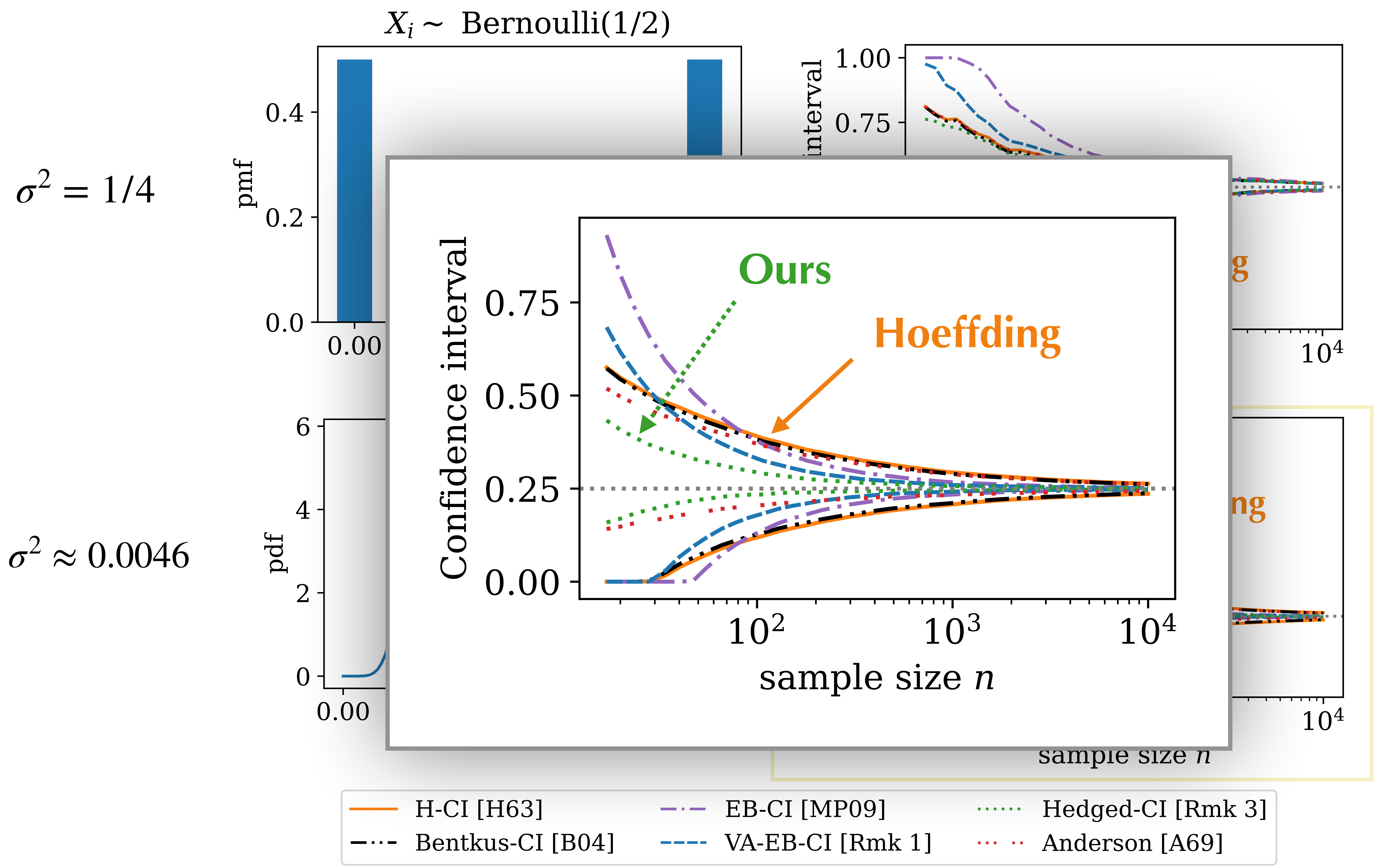


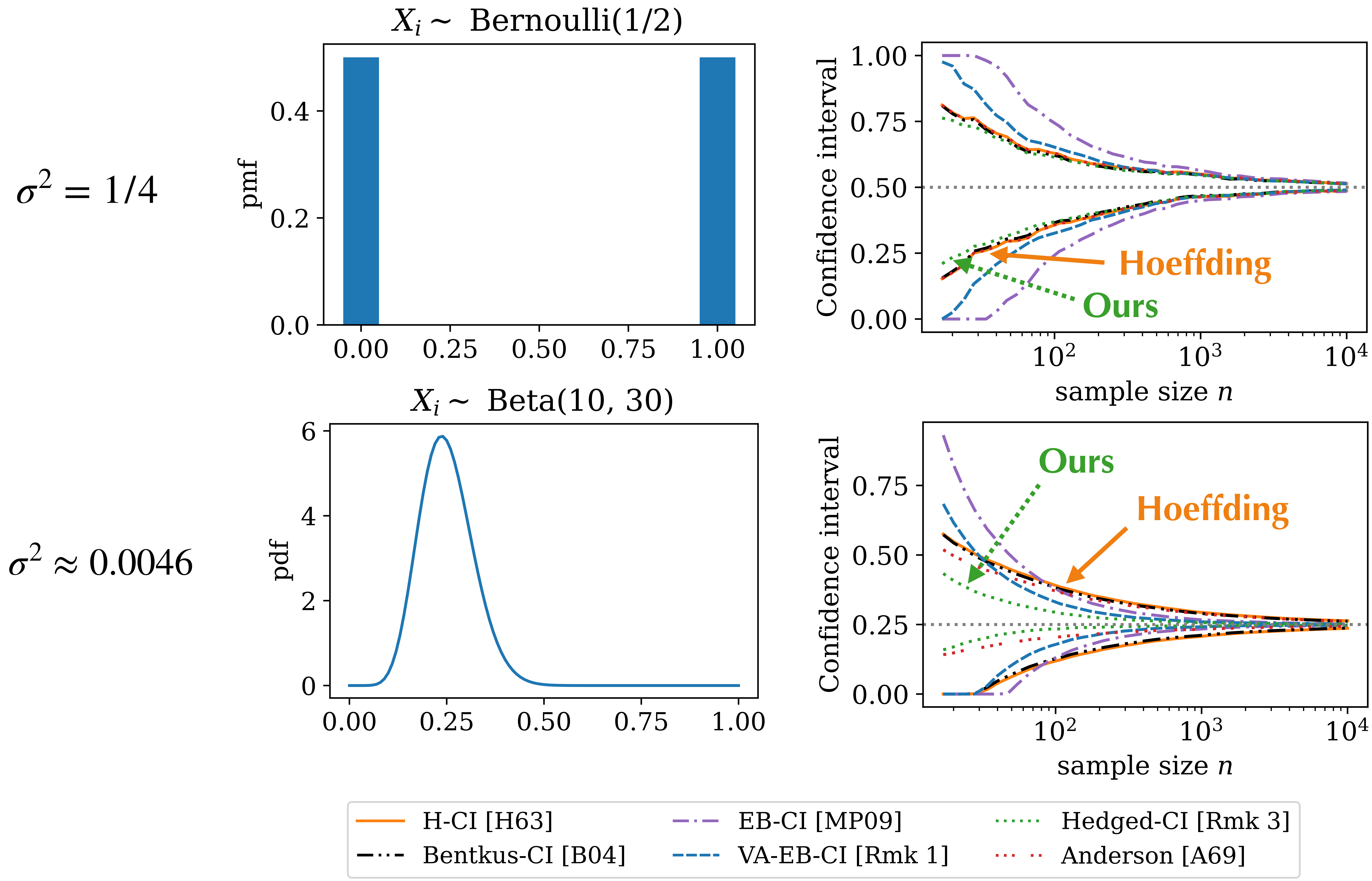
$X_i \sim$ Bernoulli(1/2)
$\sigma^2 = 1/4$
pmf
0.0
0.2
0.4
0.00
0.25
0.50
0.75
1.00
Confidence interval
sample size n
10^2
10^3
10^4
Hoeffding
Ours
$X_i \sim$ Beta(10, 30)
$\sigma^2 \approx 0.0046$
pdf
0
2
4
6
Hoeffding
H-CI [H63]
Bentkus-CI [B04]
EB-CI [MP09]
VA-EB-CI [Rmk 1]
Hedged-CI [Rmk 3]
Anderson [A69]


$\sigma^2 = 1/4$
$\sigma^2 \approx 0.0046$
$X_i \sim$ Bernoulli(1/2)
pmf
$X_i \sim$ Beta(10, 30)
pdf
Confidence interval
sample size n
Hoeffding
Ours
H-CI [H63]
Bentkus-CI [B04]
EB-CI [MP09]
VA-EB-CI [Rmk 1]
Hedged-CI [Rmk 3]
Anderson [A69]



$\sigma^2 = 1/4$
$X_i \sim$ Bernoulli(1/2)
pmf
0.0
0.2
0.4
0.00
0.25
0.50
0.75
1.00
Confidence interval
0.00
0.25
0.50
0.75
1.00
10^2
10^3
10^4
sample size n
Hoeffding
Ours
$\sigma^2 \approx 0.0046$
$X_i \sim$ Beta(10, 30)
pdf
0
2
4
6
0.00
0.25
0.50
0.75
1.00
Confidence interval
0.00
0.25
0.50
0.75
10^2
10^3
10^4
sample size n
Ours
Hoeffding
H-CI [H63]
Bentkus-CI [B04]
EB-CI [MP09]
VA-EB-CI [Rmk 1]
Hedged-CI [Rmk 3]
Anderson [A69]


$X_i \sim$ Bernoulli(1/2)
$\sigma^2 = 1/4$
$\sigma^2 \approx 0.0046$
pmf
pdf
Ours
Hoeffding
Confidence interval
sample size n
H-CI [H63]
Bentkus-CI [B04]
EB-CI [MP09]
VA-EB-CI [Rmk 1]
Hedged-CI [Rmk 3]
Anderson [A69]



$\sigma^2 = 1/4$
$X_i \sim$ Bernoulli(1/2)
pmf
Confidence interval
sample size n
Hoeffding
Ours
$\sigma^2 \approx 0.0046$
$X_i \sim$ Beta(10, 30)
pdf
H-CI [H63]
Bentkus-CI [B04]
EB-CI [MP09]
VA-EB-CI [Rmk 1]
Hedged-CI [Rmk 3]
Anderson [A69]

Formal setup:

Formal setup:

1. Observe $X_1, X_2, X_3, \ldots$

Formal setup:

1. Observe $X_1, X_2, X_3, \dots$
2. $X_t \in [0,1]$ almost surely.

Formal setup:

1. Observe $X_1, X_2, X_3, \dots$
2. $X_t \in [0,1]$ almost surely.
3. $\mathbb{E}(X_t \mid X_1^{t-1}) = \mu$.

$X_1^{t-1} \equiv (X_1, \dots, X_{t-1})$

Formal setup:

1. Observe $X_1, X_2, X_3, \ldots$
2. $X_t \in [0,1]$ almost surely.
3. $\mathbb{E}(X_t \mid X_1^{t-1}) = \mu$.

$X_1^{t-1} \equiv (X_1, \ldots, X_{t-1})$

Familiar special case: $X_1, X_2, \ldots \overset{iid}{\sim} \mathbb{P}$, with $\mathbb{E}_{\mathbb{P}}(X_1) = \mu$.

Consider a “candidate mean” $m \in [0,1]$

Consider a “candidate mean” $m \in [0,1]$

$K_0 \leftarrow \$1$

Consider a "candidate mean" $m \in [0,1]$

$K_0 \leftarrow \$1$

For $t = 1,2,3,\ldots$:

Consider a "candidate mean" $m \in [0,1]$

$K_0 \leftarrow \$1$

For $t = 1,2,3,\ldots$:

Gambler chooses bet $\lambda_t \in (-1/(1-m),\ 1/m)$

Consider a "candidate mean" $m \in [0,1]$

$K_0 \leftarrow \$1$

For $t = 1,2,3,\ldots$:

Gambler chooses bet $\lambda_t \in (-1/(1-m),\ 1/m)$

Observe X_t

Consider a "candidate mean" $m \in [0,1]$

$K_0 \leftarrow \$1$

For $t = 1,2,3,\ldots$:

Gambler chooses bet $\lambda_t \in (-1/(1-m),\ 1/m)$

Observe X_t

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

EndFor

Consider a "candidate mean" $m \in [0,1]$

$K_0 \leftarrow \$1$

For $t = 1,2,3,\ldots$:

Gambler chooses bet $\lambda_t \in (-1/(1-m),\ 1/m)$

Observe X_t

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

EndFor

What if

$\mathbb{E}(X_t) \equiv \mu \gg m$?

or $\mu \ll m$?

<u>or $\mu = m$?</u>

The "capital process" $(K_t(\mu))_{t=0}^{\infty}$

is a nonnegative martingale starting at one.

The "capital process" $(K_t(\mu))_{t=0}^{\infty}$

is a nonnegative martingale starting at one.

$$\mathbb{E}(K_t(\mu) \mid X_1^{t-1}) = K_{t-1}(\mu) + K_{t-1}(\mu) \cdot \mathbb{E}(\lambda_t(X_t - \mu) \mid X_1^{t-1})$$

The "capital process" $(K_t(\mu))_{t=0}^{\infty}$

is a nonnegative martingale starting at one.

$$\mathbb{E}(K_t(\mu) \mid X_1^{t-1}) = K_{t-1}(\mu) + K_{t-1}(\mu) \cdot \underbrace{\mathbb{E}(\lambda_t(X_t - \mu) \mid X_1^{t-1})}$$

The "capital process" $(K_t(\mu))_{t=0}^{\infty}$

is a nonnegative martingale starting at one.

$$\mathbb{E}(K_t(\mu) \mid X_1^{t-1}) = K_{t-1}(\mu) + K_{t-1}(\mu) \cdot \underbrace{\mathbb{E}(\lambda_t(X_t - \mu) \mid X_1^{t-1})}_{= \lambda_t \left(\mathbb{E}(X_t \mid X_1^{t-1}) - \mu\right)}$$

The "capital process" $(K_t(\mu))_{t=0}^{\infty}$

is a nonnegative martingale starting at one.

$$\mathbb{E}(K_t(\mu) \mid X_1^{t-1}) = K_{t-1}(\mu) + K_{t-1}(\mu) \cdot \underbrace{\mathbb{E}(\lambda_t(X_t - \mu) \mid X_1^{t-1})}_{\substack{= \lambda_t \left(\mathbb{E}(X_t \mid X_1^{t-1}) - \mu\right) \\ = 0}}$$

The "capital process" $(K_t(\mu))_{t=0}^{\infty}$

is a nonnegative martingale starting at one.

$$\mathbb{E}(K_t(\mu) \mid X_1^{t-1}) = K_{t-1}(\mu) + K_{t-1}(\mu) \cdot \underbrace{\mathbb{E}(\lambda_t(X_t - \mu) \mid X_1^{t-1})}_{\substack{= \lambda_t \left(\mathbb{E}(X_t \mid X_1^{t-1}) - \mu\right) \\ = 0}}$$

$$\therefore \mathbb{E}(K_t(\mu) \mid X_1^{t-1}) = K_{t-1}(\mu)$$

The "capital process" $(K_t(\mu))_{t=0}^{\infty}$

is a nonnegative martingale starting at one.

$$\mathbb{E}(K_t(\mu) \mid X_1^{t-1}) = K_{t-1}(\mu) + K_{t-1}(\mu) \cdot \underbrace{\mathbb{E}(\lambda_t(X_t - \mu) \mid X_1^{t-1})}$$

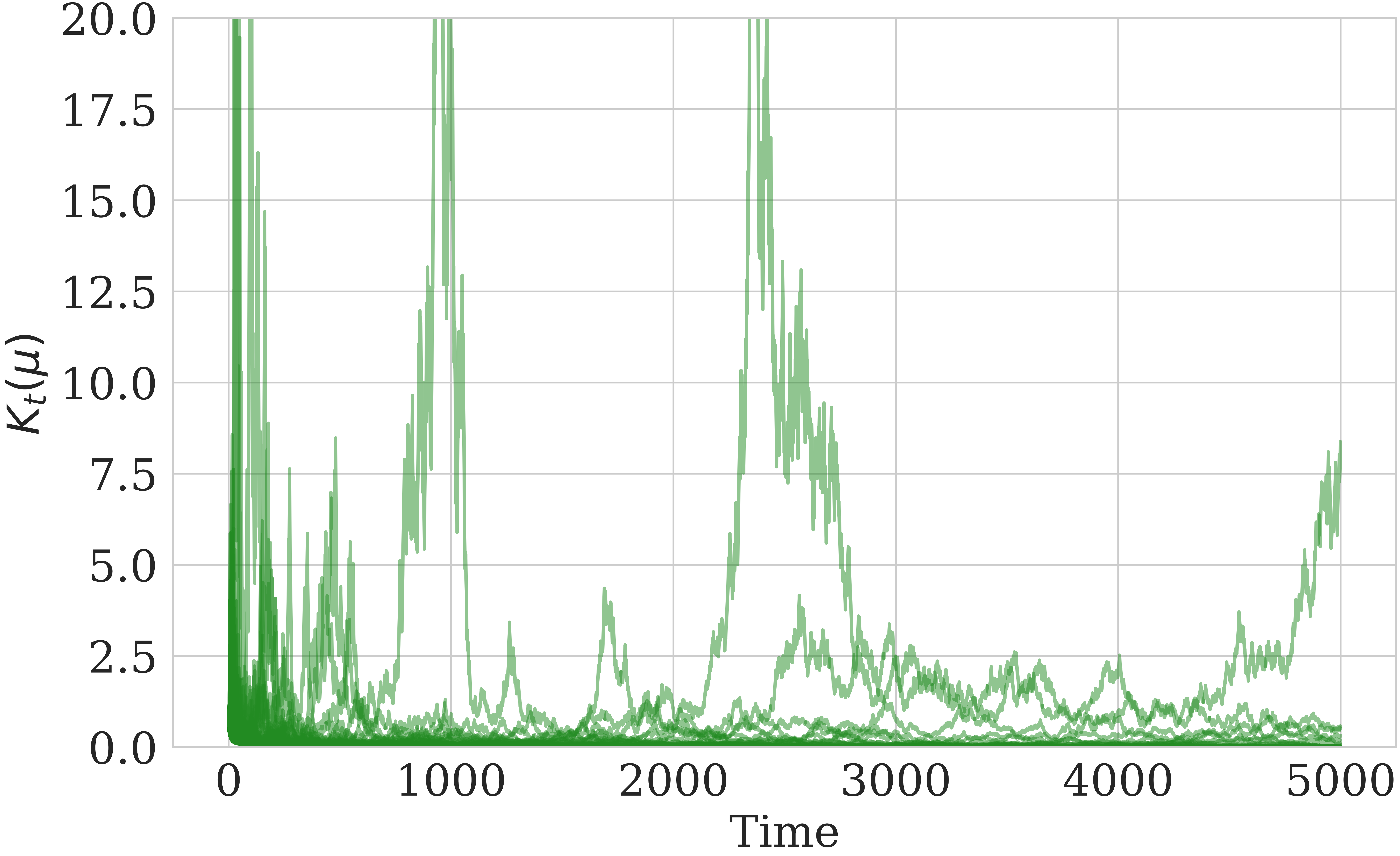
$$= \lambda_t \left(\mathbb{E}(X_t \mid X_1^{t-1}) - \mu \right)$$

$$= 0$$

$$\therefore \mathbb{E}(K_t(\mu) \mid X_1^{t-1}) = K_{t-1}(\mu)$$

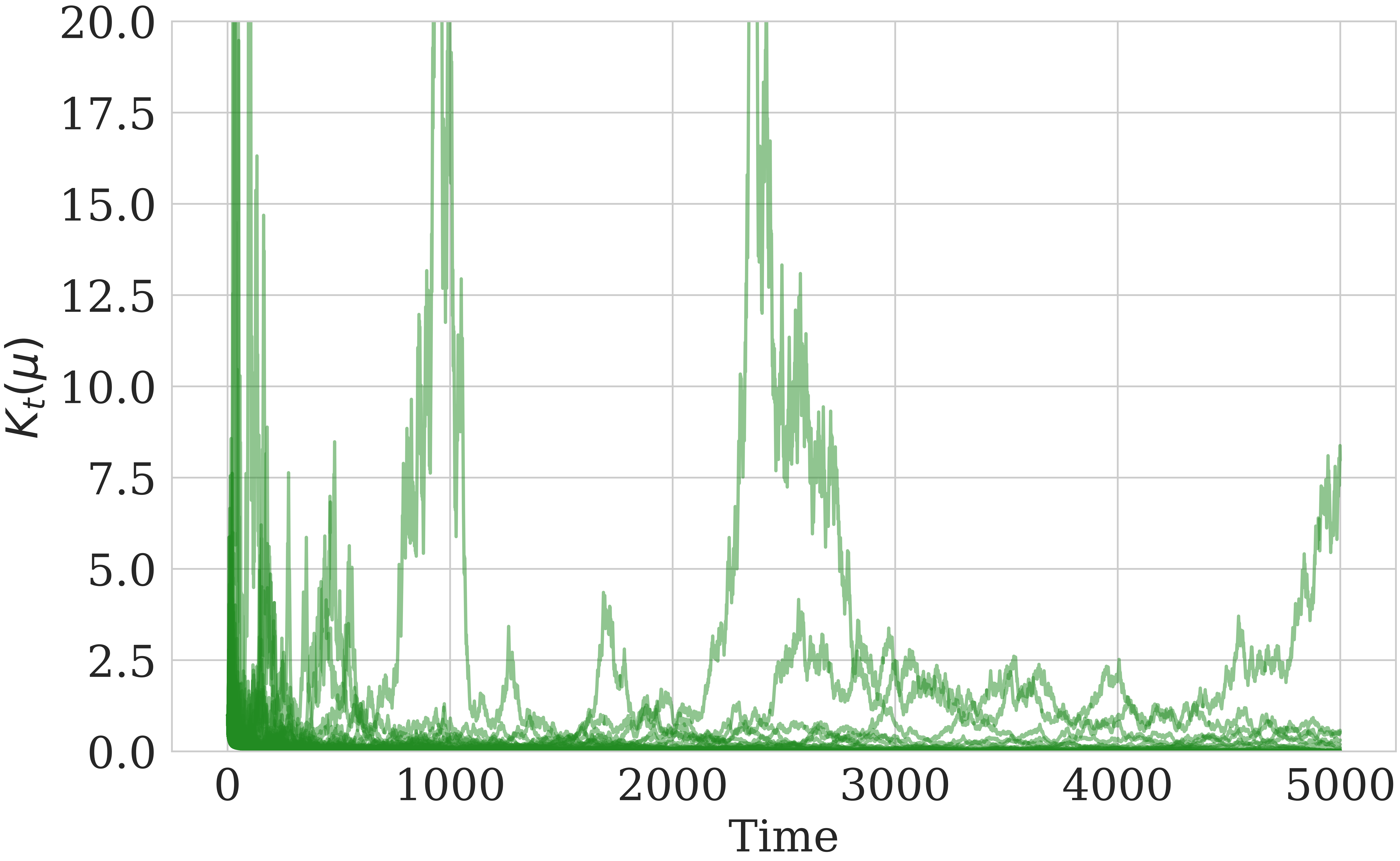
Similar ideas from Shafer & Vovk, Hendriks, Jun & Orabona, etc.

Wealth processes / nonnegative martingales

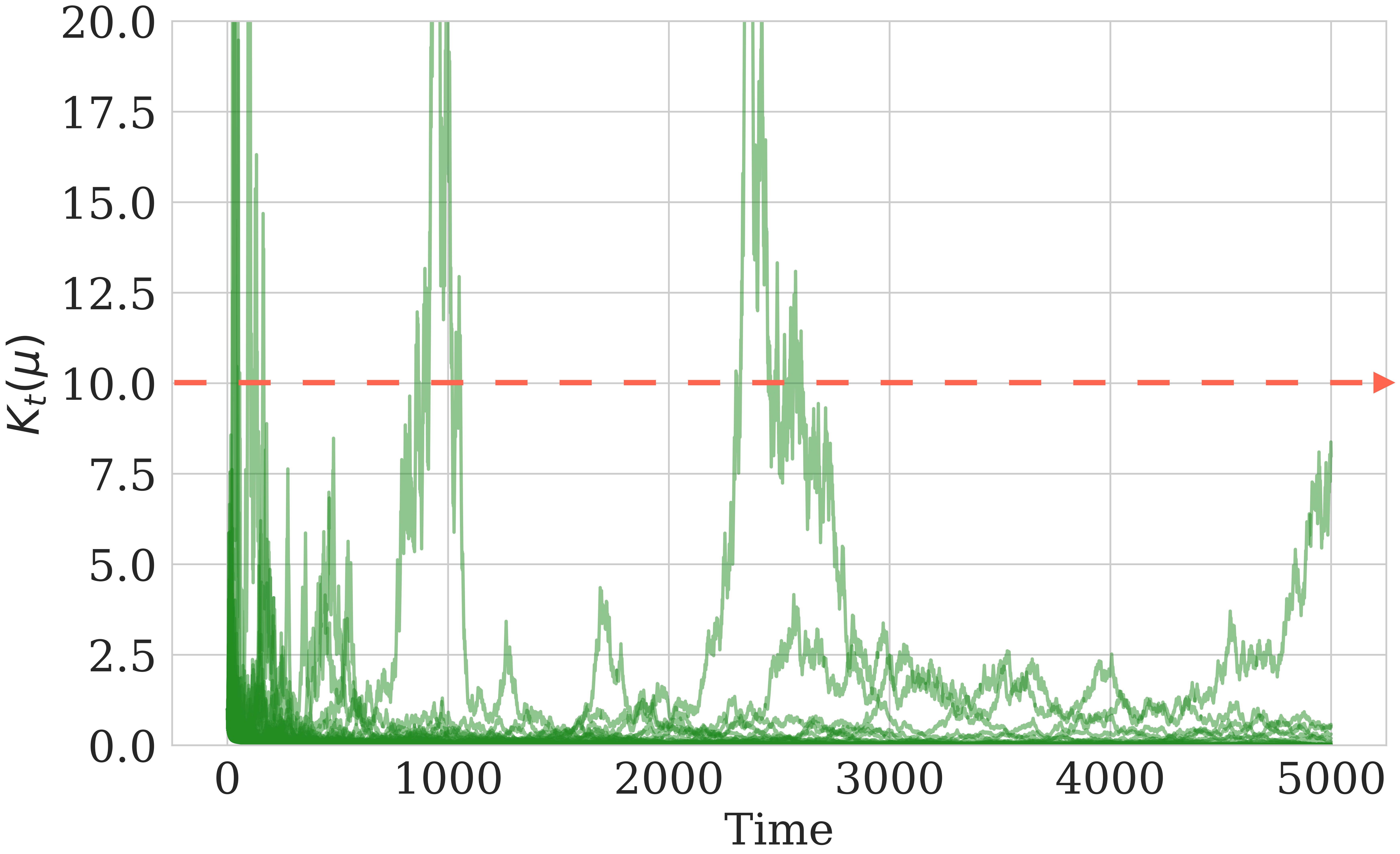


Jean Ville (1939): $\mathbb{P}\left(\exists t \geq 1 : K_t(\mu) \geq \frac{1}{\alpha}\right) \leq \alpha$

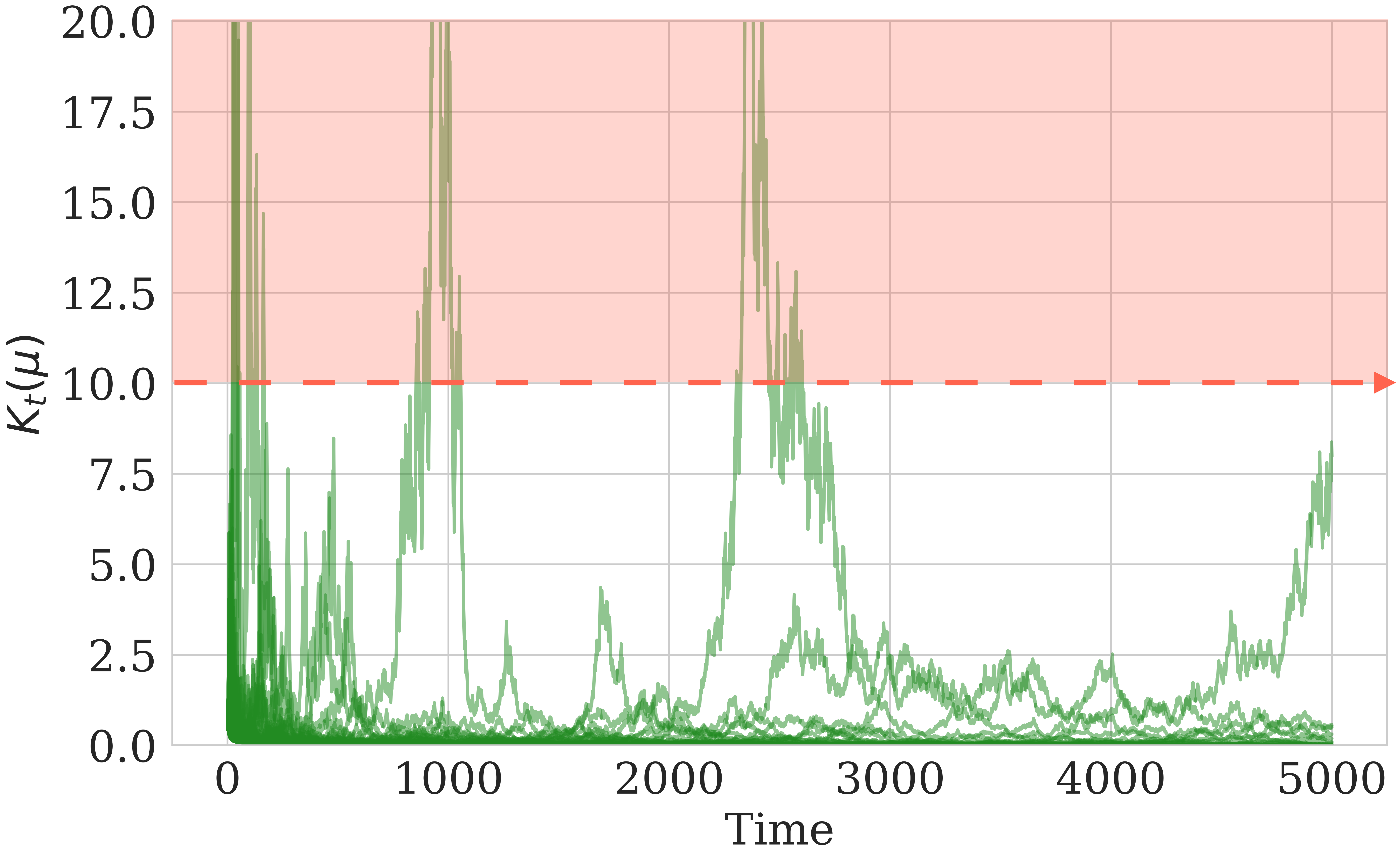
Jean Ville (1939): $\mathbb{P}\left(\exists t \geq 1 : K_t(\mu) \geq \frac{1}{\alpha}\right) \leq \alpha$



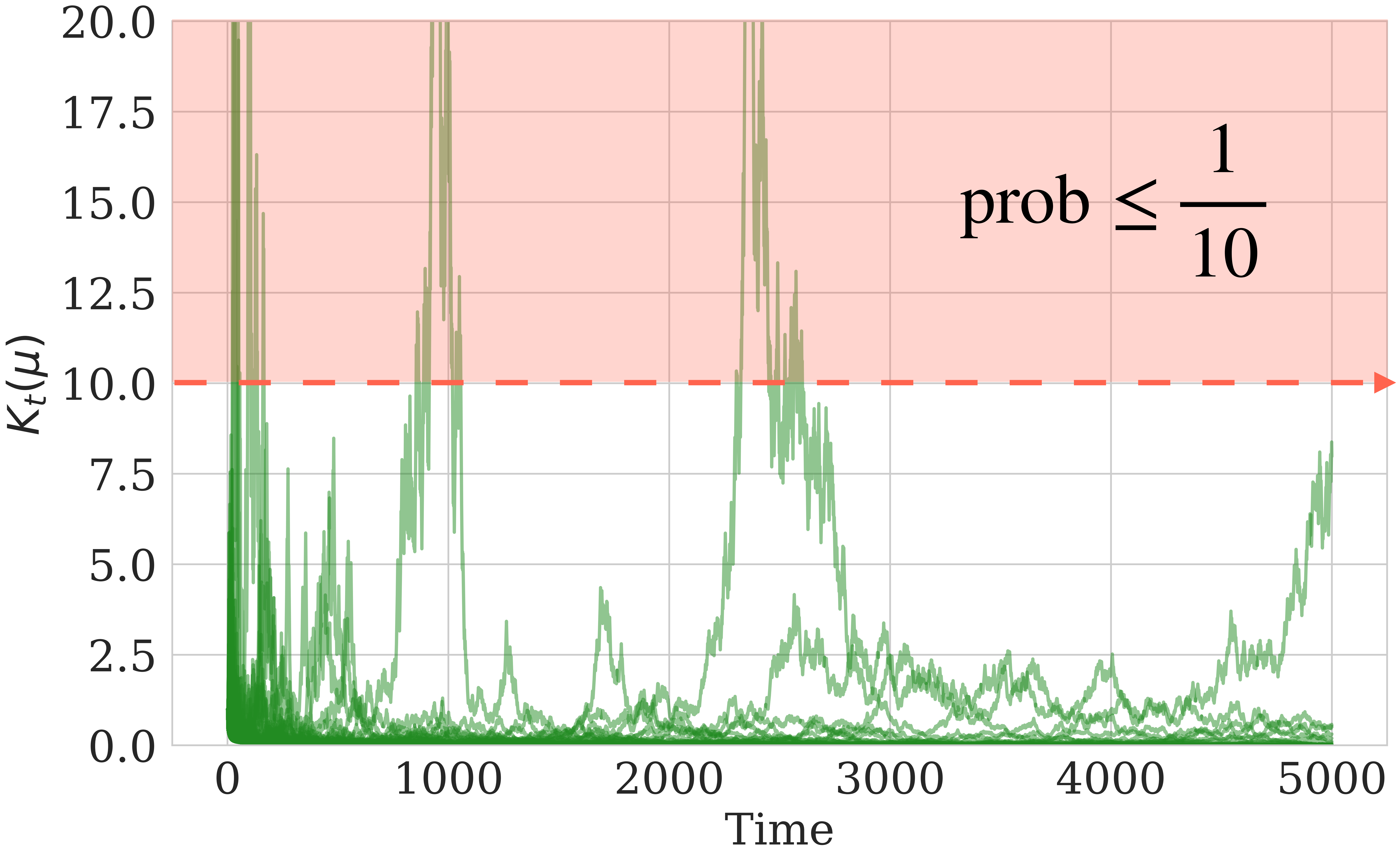
Jean Ville (1939): $\mathbb{P}\left(\exists t \geq 1 : K_t(\mu) \geq \frac{1}{\alpha}\right) \leq \alpha$



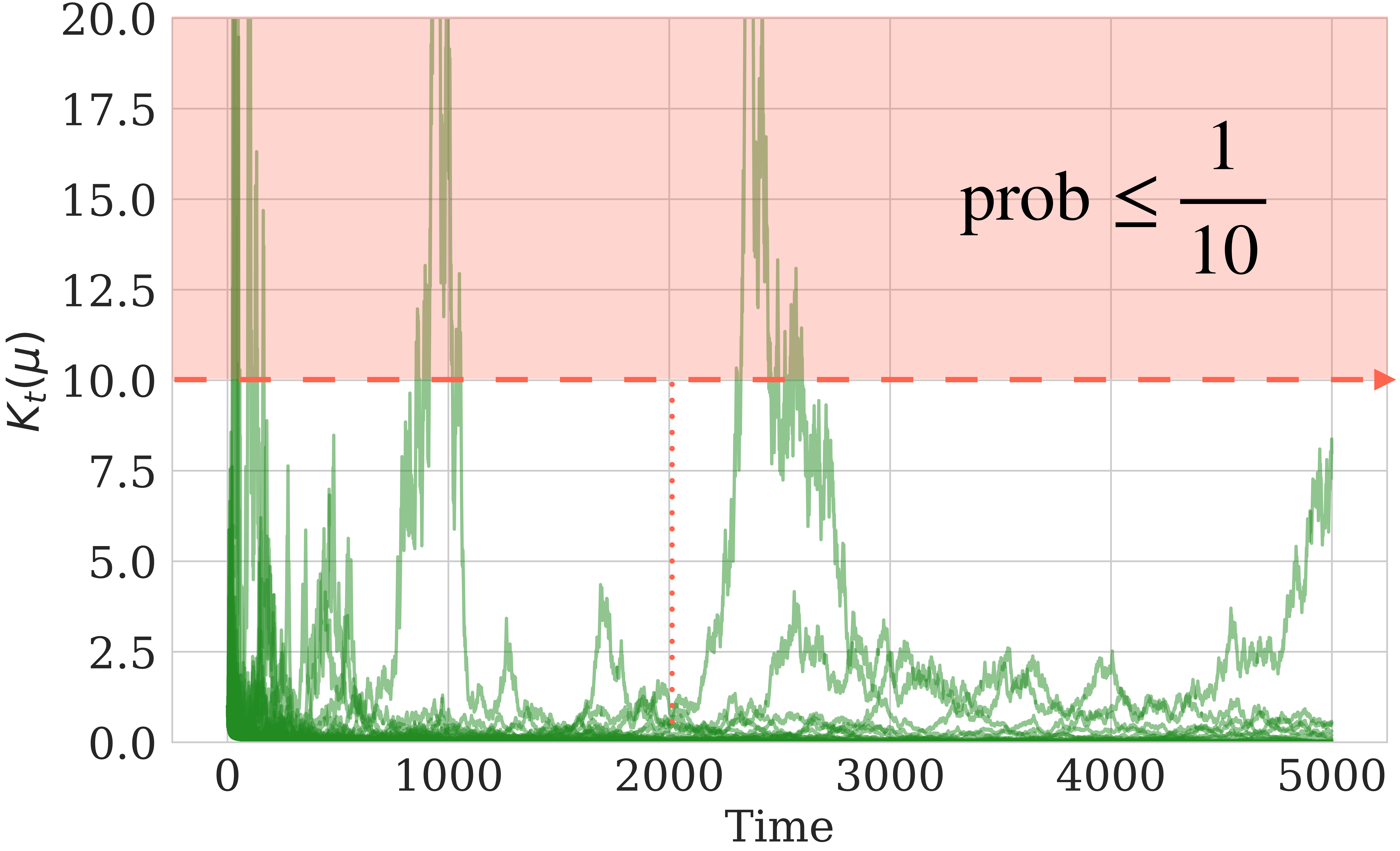
Jean Ville (1939): $\mathbb{P}\left(\exists t \geq 1 : K_t(\mu) \geq \frac{1}{\alpha}\right) \leq \alpha$



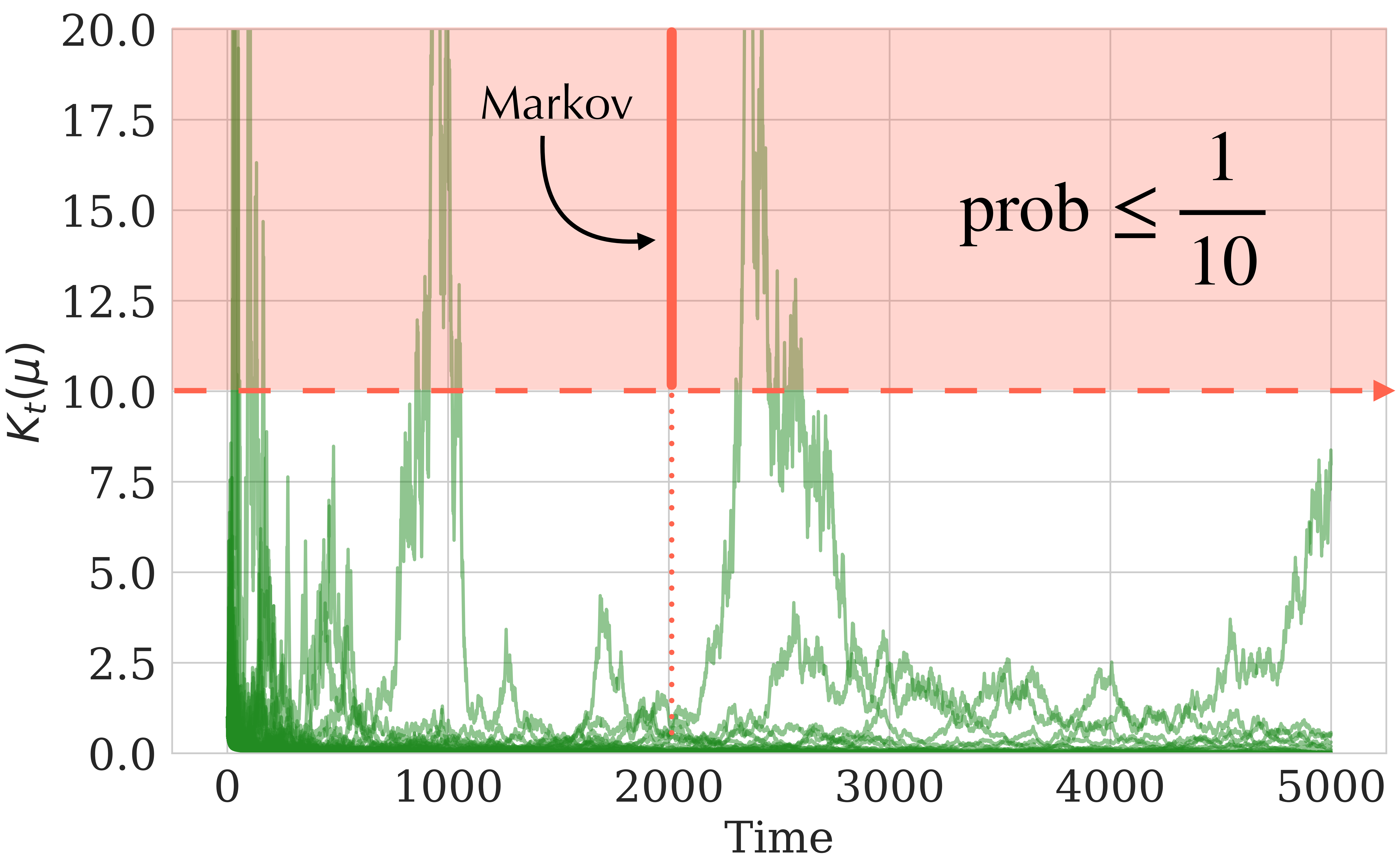
Jean Ville (1939): $\mathbb{P}\left(\exists t \geq 1 : K_t(\mu) \geq \frac{1}{\alpha}\right) \leq \alpha$



Jean Ville (1939): $\mathbb{P}\left(\exists t \geq 1 : K_t(\mu) \geq \frac{1}{\alpha}\right) \leq \alpha$



Jean Ville (1939): $\mathbb{P}\left(\exists t \geq 1 : K_t(\mu) \geq \frac{1}{\alpha}\right) \leq \alpha$



“Invert” Ville’s inequality:

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

Robbins et al. (1960s-1970s),
Shafer & Vovk (2001–2019)
Johari et al. (2015)
Jun & Orabona (2019–),
Howard et al. (2019–).
Grunwald (2019–)

“Invert” Ville’s inequality:

Robbins et al. (1960s-1970s),
Shafer & Vovk (2001–2019)
Johari et al. (2015)
Jun & Orabona (2019–),
Howard et al. (2019–).
Grunwald (2019–)

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

“The set of m for which the gambler didn’t make much money”.

“Invert” Ville’s inequality:

Robbins et al. (1960s-1970s),
Shafer & Vovk (2001–2019)
Johari et al. (2015)
Jun & Orabona (2019–),
Howard et al. (2019–).
Grunwald (2019–)

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

“The set of m for which the gambler didn’t make much money”.

$$\mathbb{P}\left(\exists t \geq 1 : \mu \notin C_t \right) \leq \alpha\,.$$

“Invert” Ville’s inequality:

Robbins et al. (1960s-1970s),
Shafer & Vovk (2001–2019)
Johari et al. (2015)
Jun & Orabona (2019–),
Howard et al. (2019–).
Grunwald (2019–)

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

“The set of m for which the gambler didn’t make much money”.

$$\mathbb{P}\left(\exists t \geq 1 : \mu \notin C_t \right) \leq \alpha .$$

$(C_t)_{t=1}^{\infty}$ forms a $(1-\alpha)$-***confidence sequence.***

Detour: confidence sequences

Detour: confidence sequences

Herbert Robbins, 1960s/70s

+ Siegmund, Darling, & Lai

Confidence sequence

Confidence interval

Confidence sequence

Confidence interval

$$\mathbb{P}\left(\exists t \geq 1 : \mu \notin C_t\right) \leq \alpha$$

Confidence sequence

$$\mathbb{P}\left(\exists t \geq 1 : \mu \notin C_t\right) \leq \alpha$$

Confidence interval

$$\forall n, \mathbb{P}\left(\mu \notin C_n\right) \leq \alpha$$

Confidence sequence

$$\mathbb{P}\left(\exists t \geq 1 : \mu \notin C_t\right) \leq \alpha$$

$$\mathbb{P}\left(\forall t \geq 1, \mu \in C_t\right) \geq 1 - \alpha$$

Confidence interval

$$\forall n, \mathbb{P}\left(\mu \notin C_n\right) \leq \alpha$$

Confidence sequence

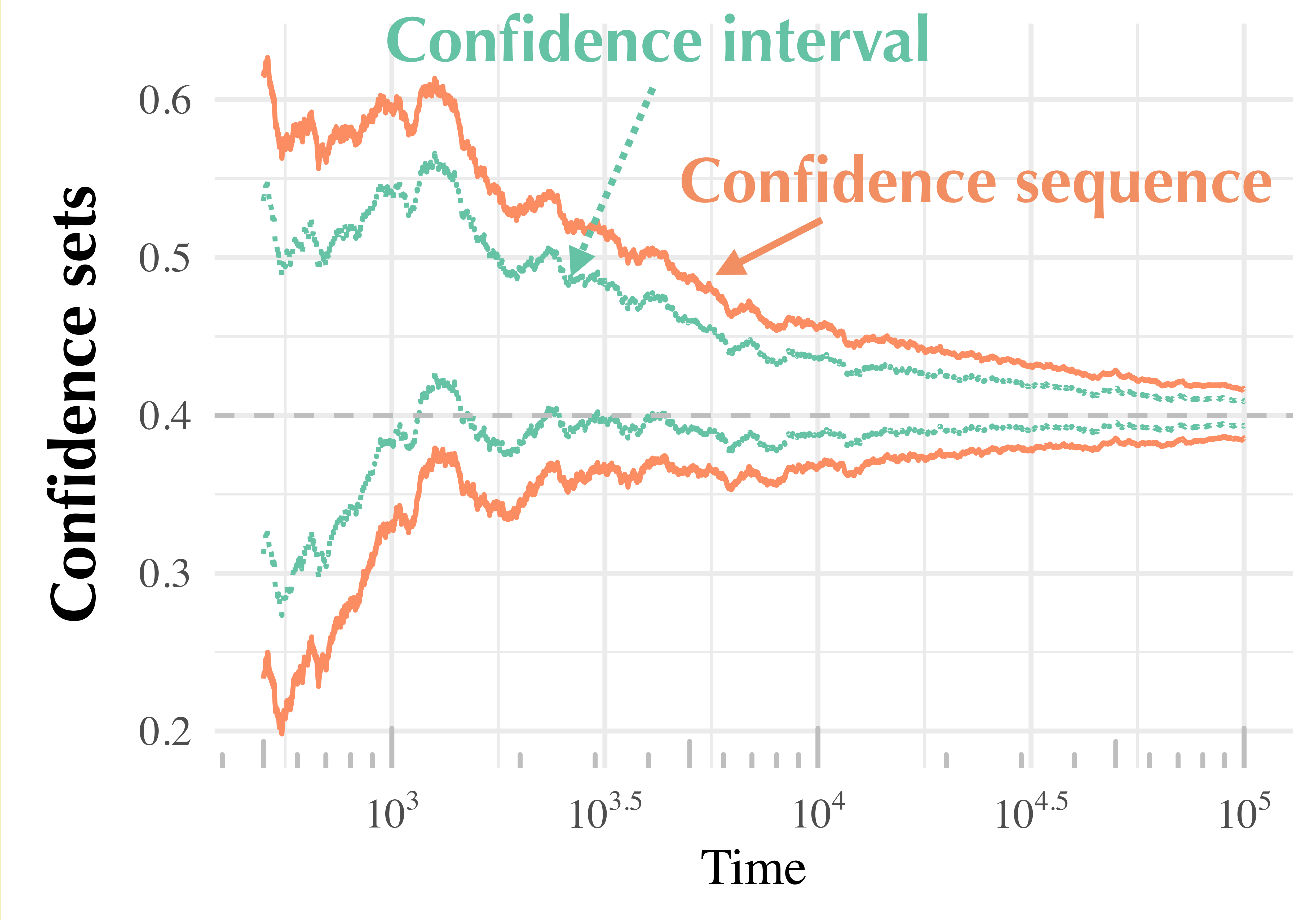
$$\mathbb{P}\left(\exists t \geq 1 : \mu \notin C_t\right) \leq \alpha$$

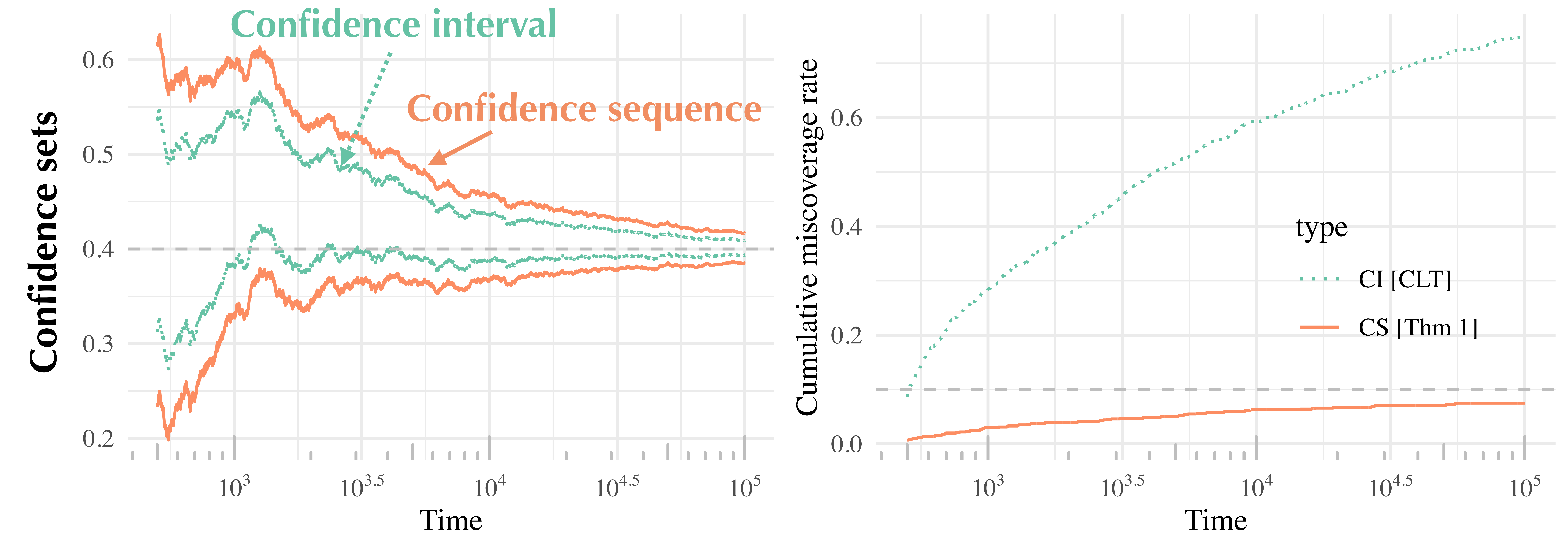
$$\mathbb{P}\left(\forall t \geq 1, \mu \in C_t\right) \geq 1 - \alpha$$

Confidence interval

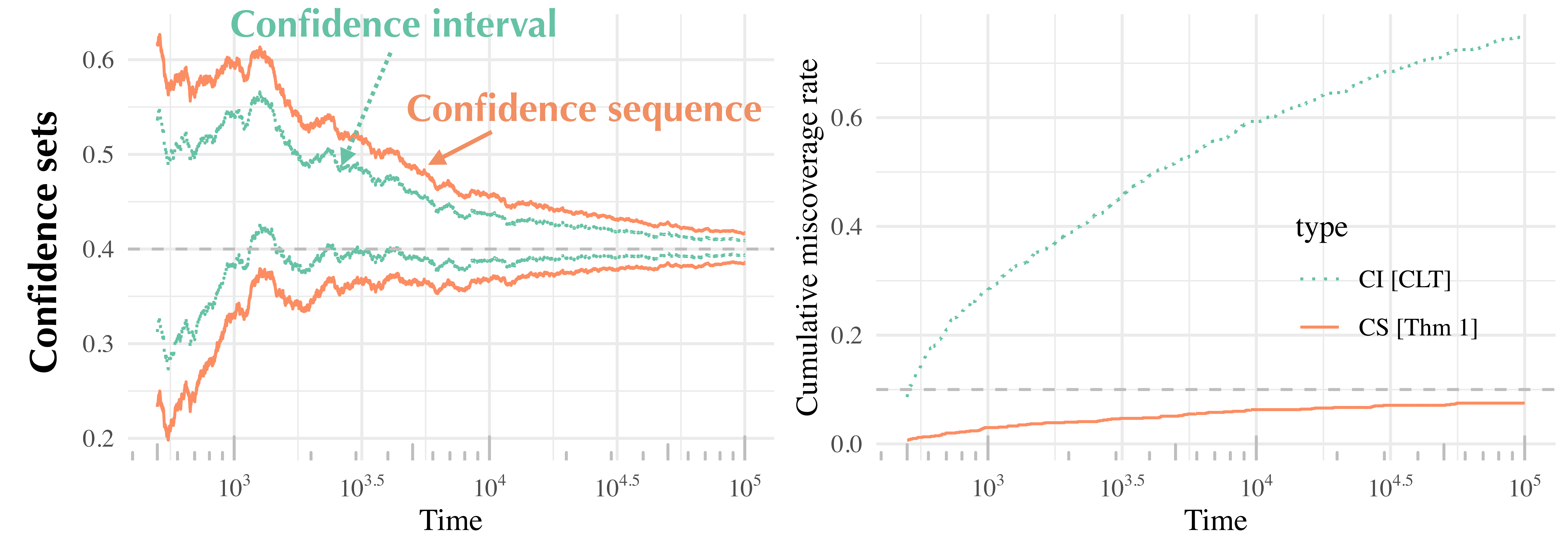
$$\forall n, \mathbb{P}\left(\mu \notin C_n\right) \leq \alpha$$

$$\forall n, \mathbb{P}\left(\mu \in C_n\right) \geq 1 - \alpha$$

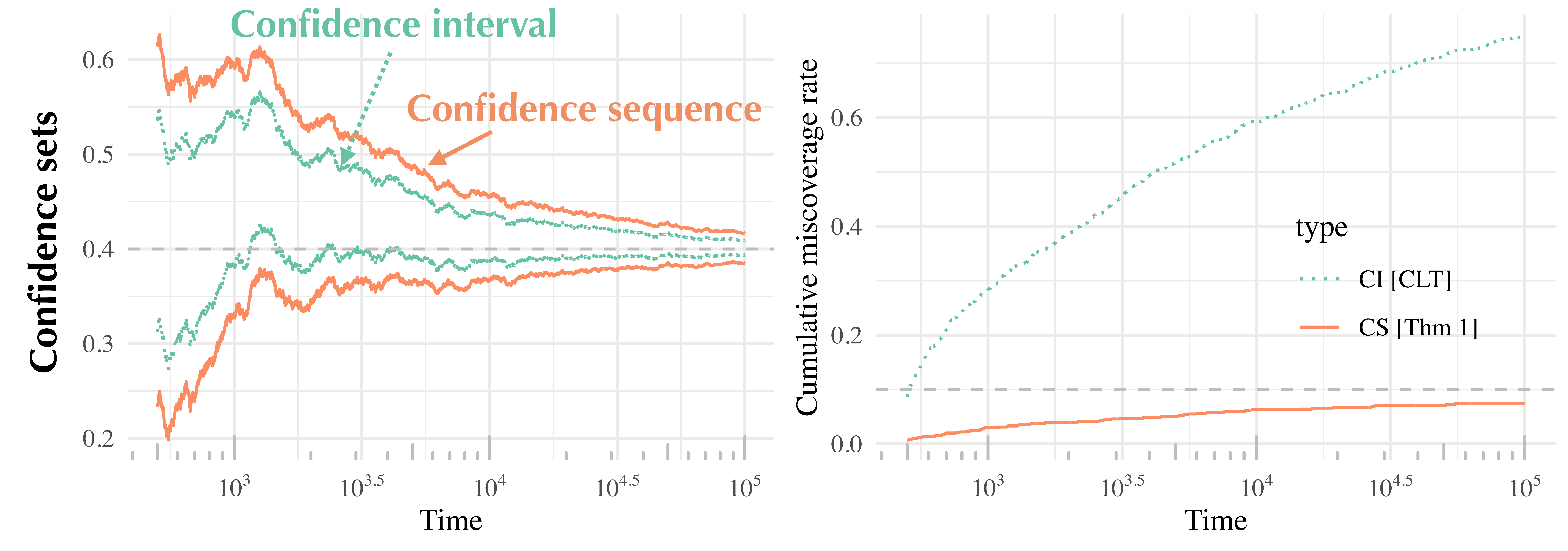

Confidence interval
Confidence sequence
Confidence sets
0.6
0.5
0.4
0.3
0.2
10^3
$10^{3.5}$
10^4
$10^{4.5}$
10^5
Time



Confidence interval
Confidence sequence
Confidence sets
0.6
0.5
0.4
0.3
0.2
10^3
$10^{3.5}$
10^4
$10^{4.5}$
10^5
Time
Cumulative miscoverage rate
0.6
0.4
0.2
0.0
type
CI [CLT]
CS [Thm 1]
10^3
$10^{3.5}$
10^4
$10^{4.5}$
10^5
Time



Confidence intervals are valid *at a single sample size*.



Confidence intervals are valid *at a single sample size*.
Confidence sequences are valid *at all sample sizes simultaneously*.

Back to confidence sequences for means of bounded random variables

Our confidence sequence:

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

is valid for any λ_i but what is a smart choice?

Maximize the **G**rowth **R**ate **A**dapted to the **P**articular **A**lternative (**GRAPA**).

$$\text{Choose } \lambda_t(m) = \arg\max_{\lambda} \frac{1}{t-1} \sum_{i=1}^{t-1} \log\left(1+\lambda \cdot (X_i - m)\right)$$

Maximize the **G**rowth **R**ate **A**dapted to the **P**articular **A**lternative (**GRAPA**).

$$\text{Choose } \lambda_t(m) = \arg\max_{\lambda} \frac{1}{t-1} \sum_{i=1}^{t-1} \log\left(1+\lambda \cdot (X_i - m)\right)$$

$$\text{Or } \lambda_t(m) \approx \frac{\widehat{\mu}_{t-1} - m}{\widehat{\sigma}^2_{t-1} + (\widehat{\mu}_{t-1} - m)^2}$$

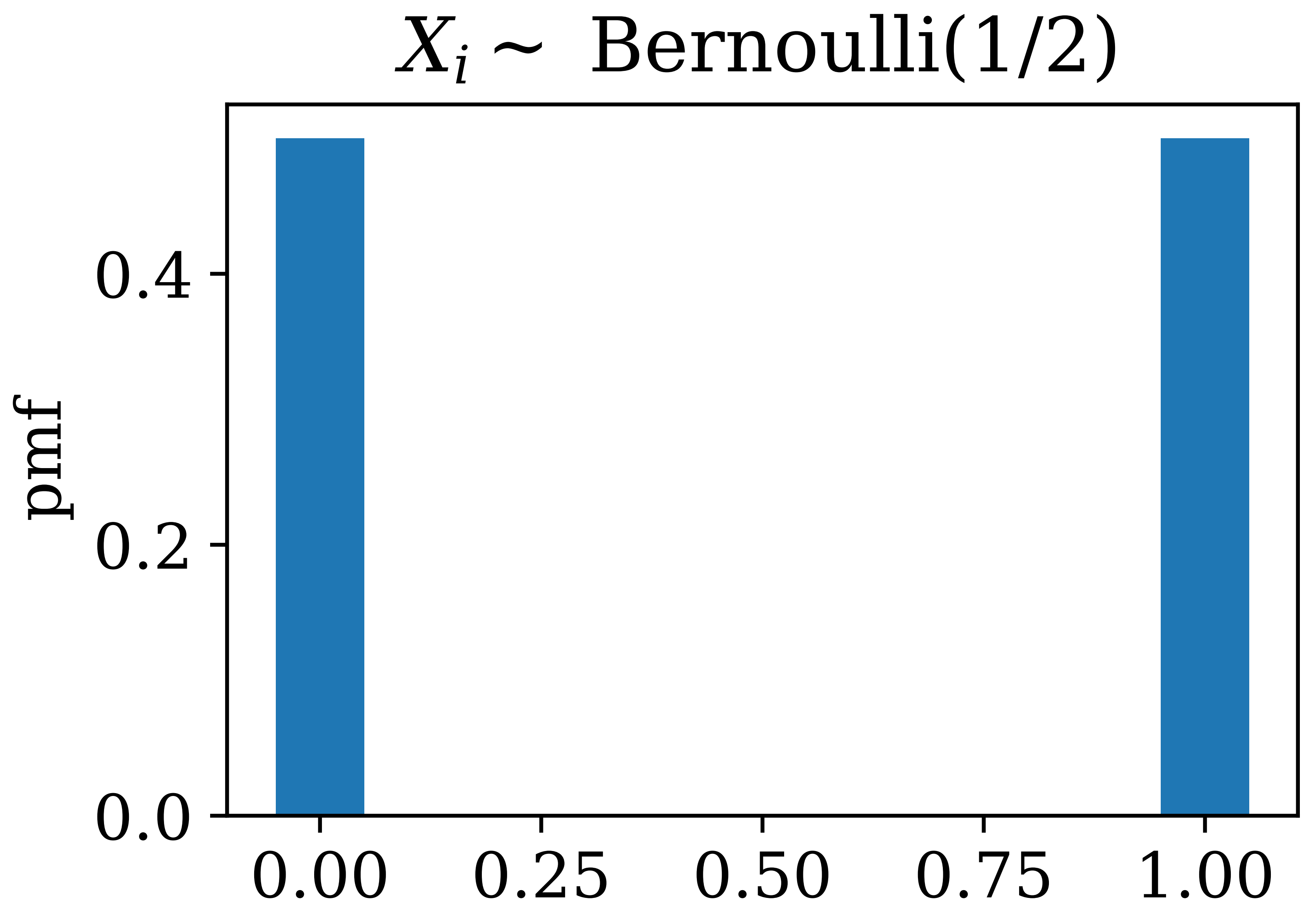
Maximize the **G**rowth **R**ate **A**dapted to the **P**articular **A**lternative (**GRAPA**).

$$\text{Choose } \lambda_t(m) = \arg\max_{\lambda} \frac{1}{t-1} \sum_{i=1}^{t-1} \log\left(1+\lambda \cdot (X_i - m)\right)$$

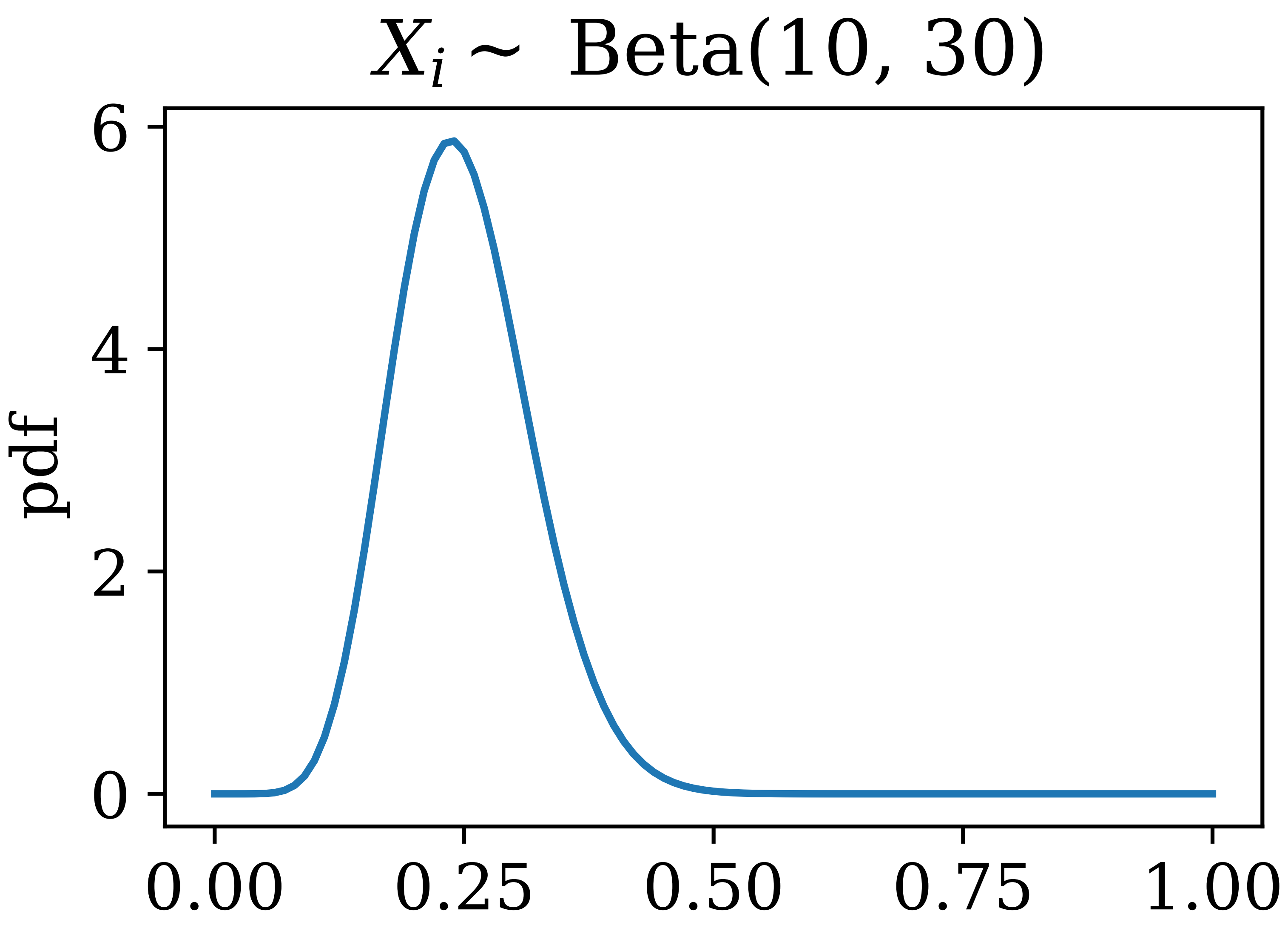
$$\text{Or } \lambda_t(m) \approx \frac{\widehat{\mu}_{t-1} - m}{\widehat{\sigma}^2_{t-1} + (\widehat{\mu}_{t-1} - m)^2}$$

$$\vdots$$

$\sigma^2 = 1/4$



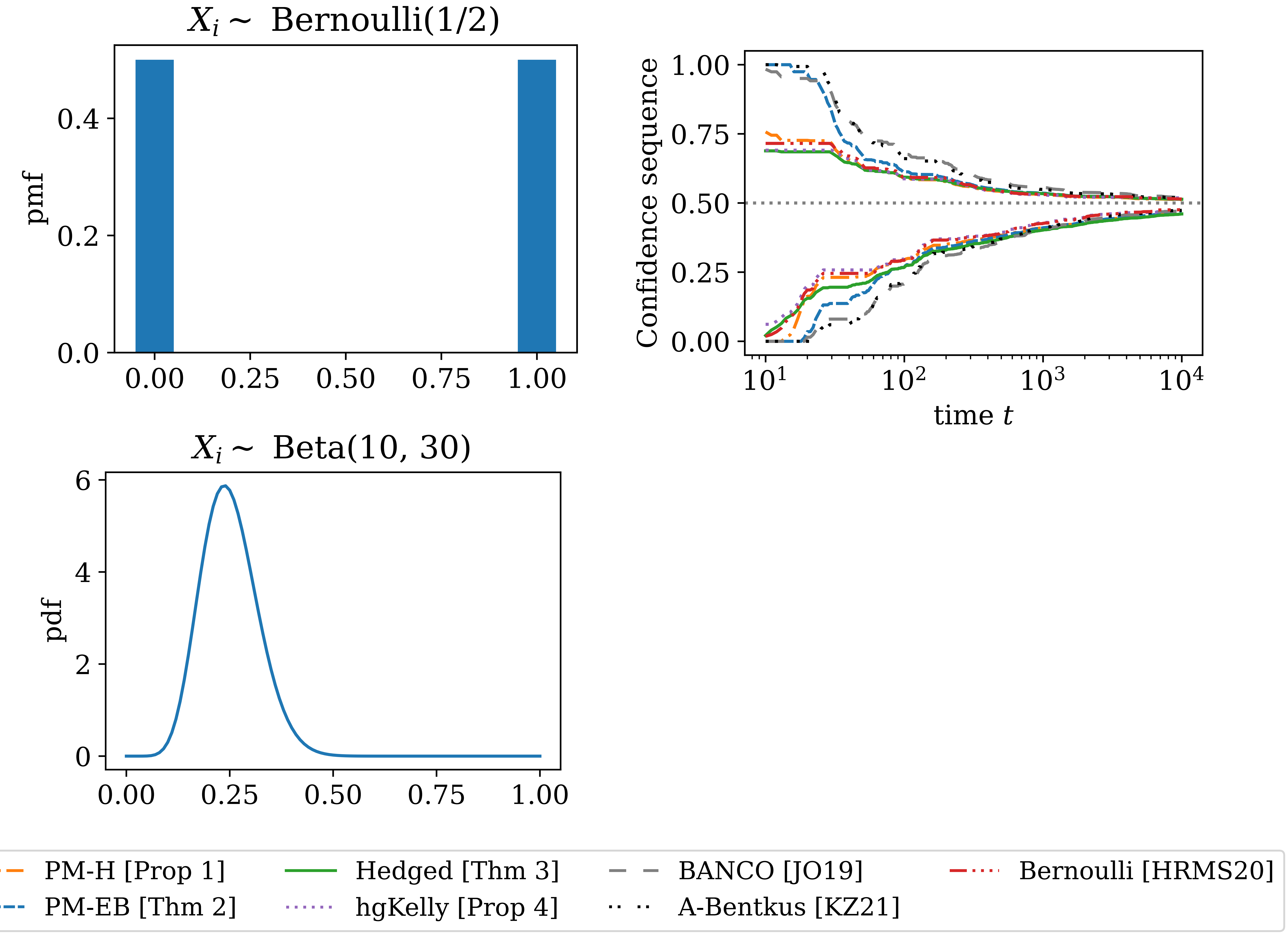
$\sigma^2 \approx 0.0046$



PM-H [Prop 1]
PM-EB [Thm 2]
Hedged [Thm 3]
hgKelly [Prop 4]
BANCO [JO19]
A-Bentkus [KZ21]
Bernoulli [HRMS20]

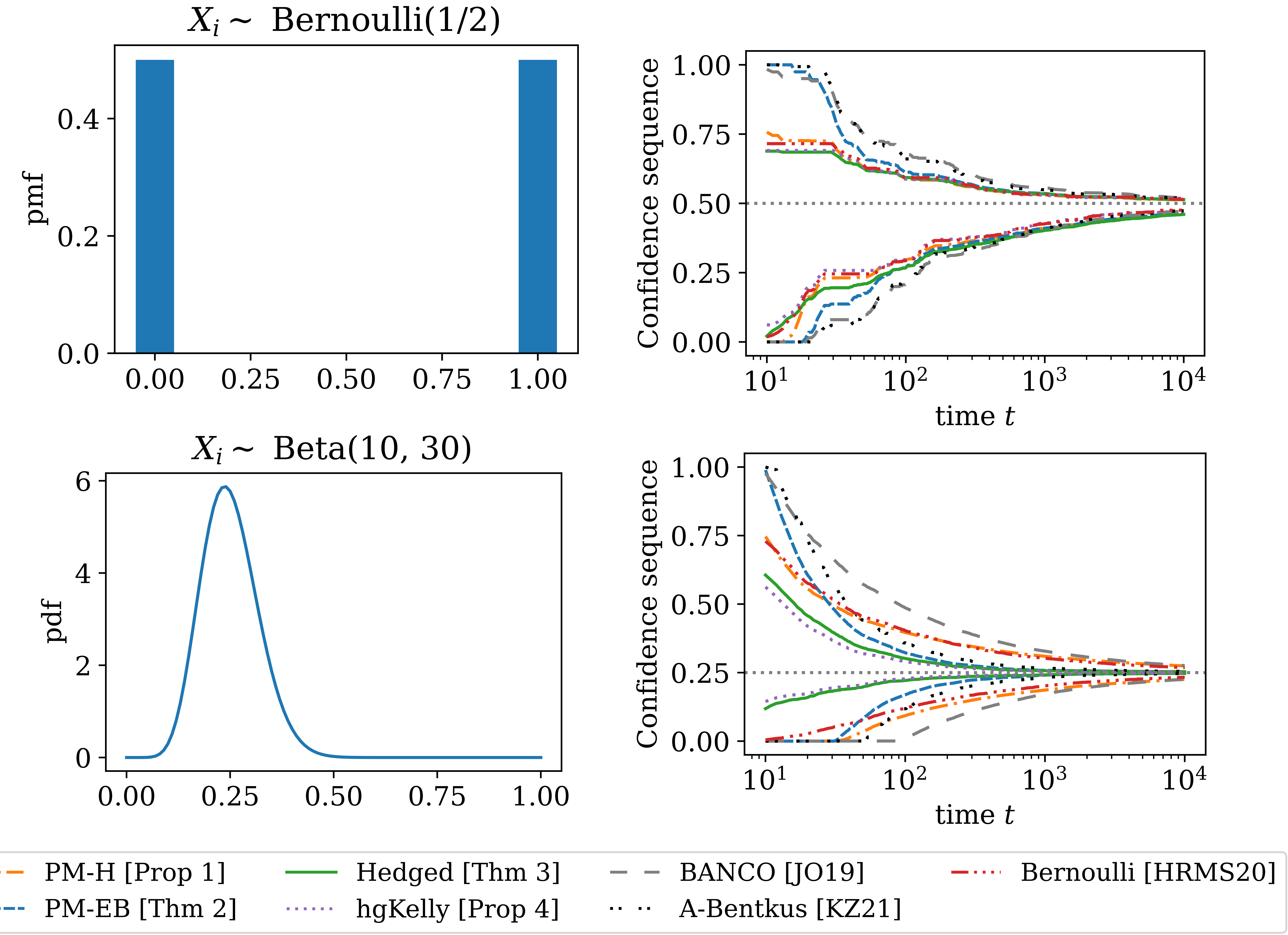
$\sigma^2 = 1/4$

$\sigma^2 \approx 0.0046$



$\sigma^2 = 1/4$

$\sigma^2 \approx 0.0046$



Contributions

Contributions

1. Developed nonparametric confidence sets for means of bounded random variables,

Contributions

1. Developed nonparametric confidence sets for means of bounded random variables,

2. Valid at arbitrary stopping times, w/ no penalties for peeking at data early.

Contributions

1. Developed nonparametric confidence sets for means of bounded random variables,

2. Valid at arbitrary stopping times, w/ no penalties for peeking at data early.

3. Substantially outperform prior work on this problem.

Contributions

1. Developed nonparametric confidence sets for means of bounded random variables,

2. Valid at arbitrary stopping times, w/ no penalties for peeking at data early.

3. Substantially outperform prior work on this problem.

+ Closed-form empirical Bernstein confidence sets*

Contributions

1. Developed nonparametric confidence sets for means of bounded random variables,

2. Valid at arbitrary stopping times, w/ no penalties for peeking at data early.

3. Substantially outperform prior work on this problem.

+ Closed-form empirical Bernstein confidence sets*
+ Extensions to sampling without replacement*

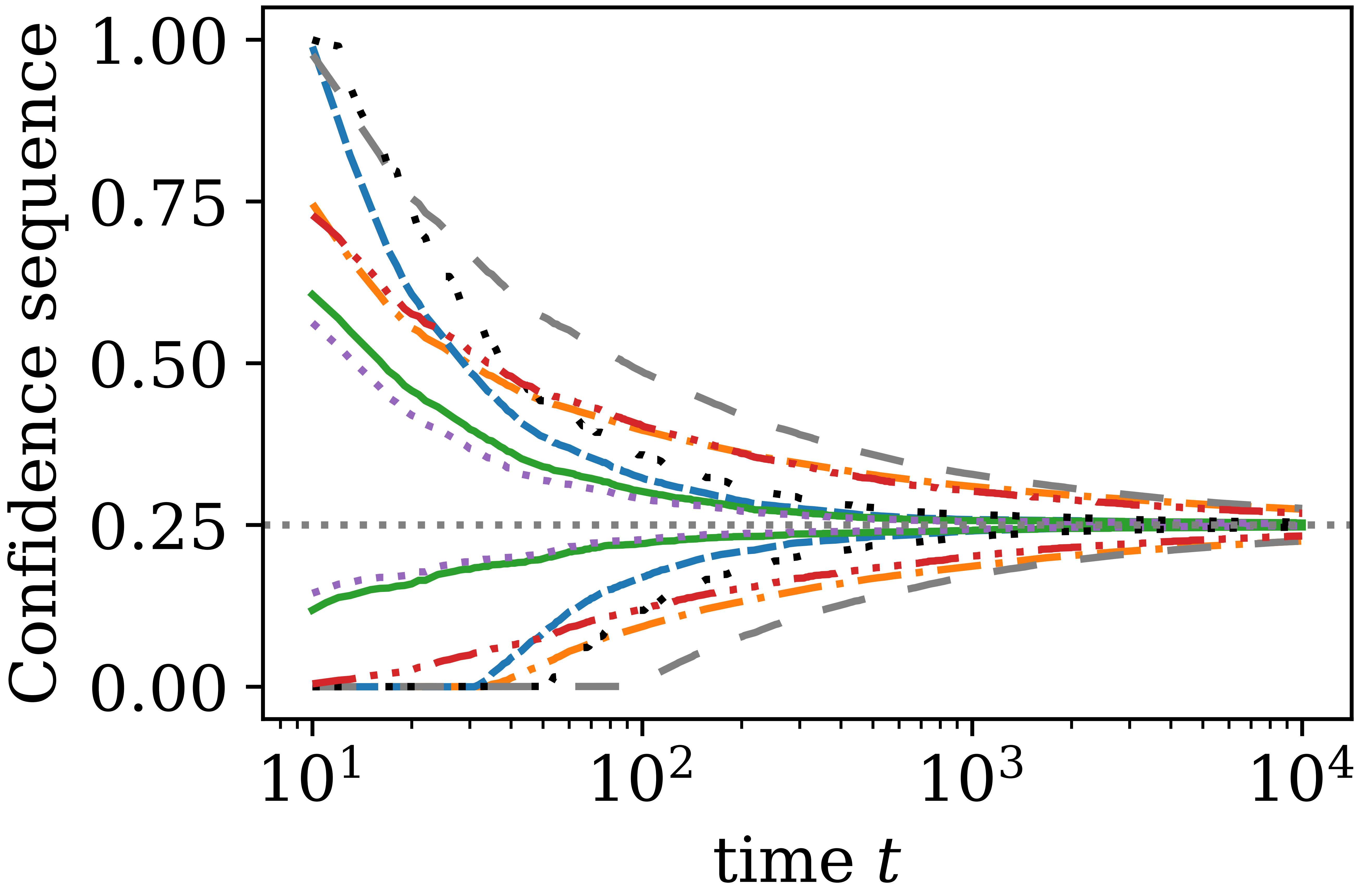
Thank you.

ian.waudbysmith.com

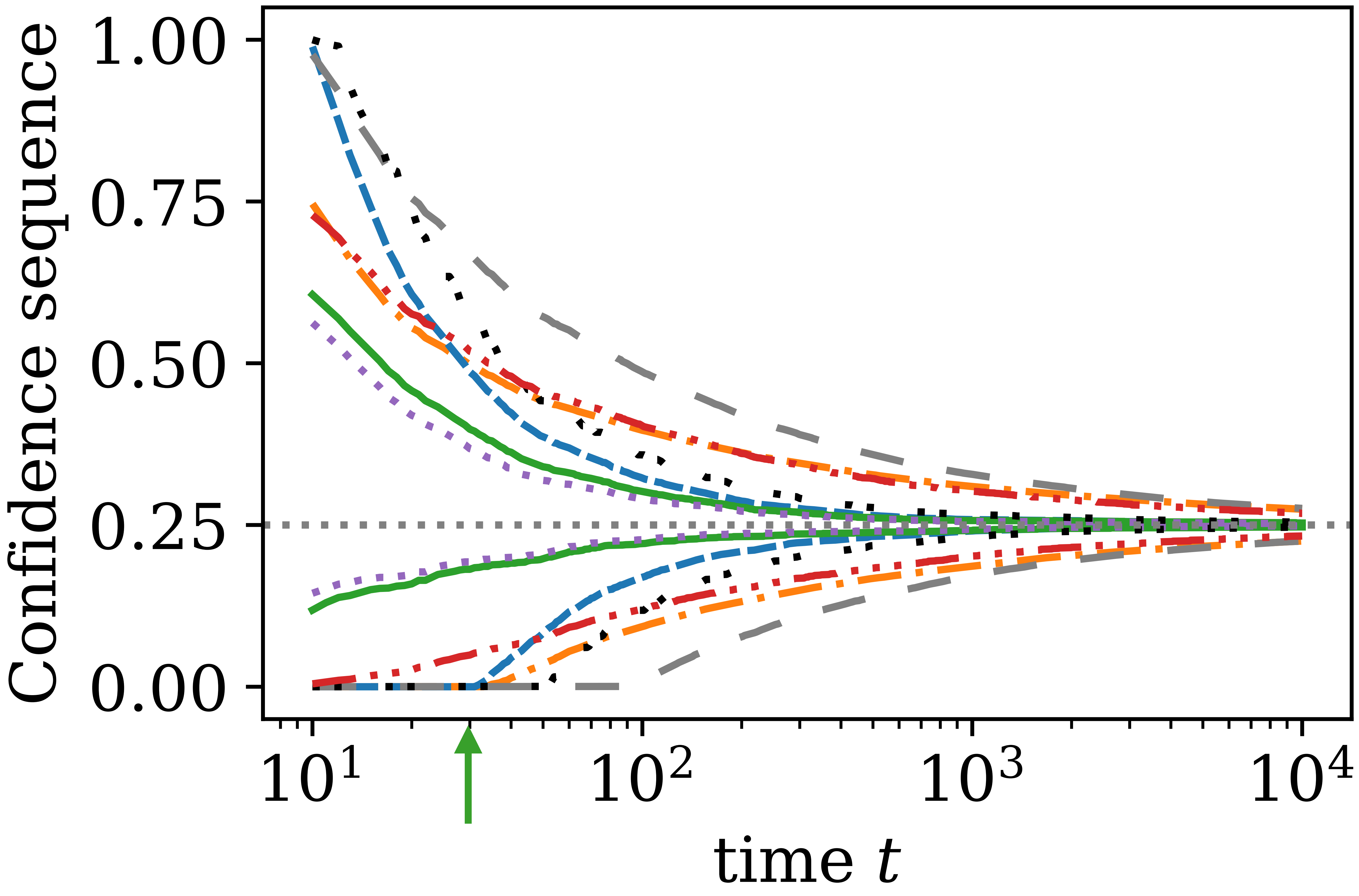
Fixed-time confidence intervals

If $(C_t)_{t=1}^{\infty}$ is a confidence sequence, then C_n is a confidence interval for any fixed n.

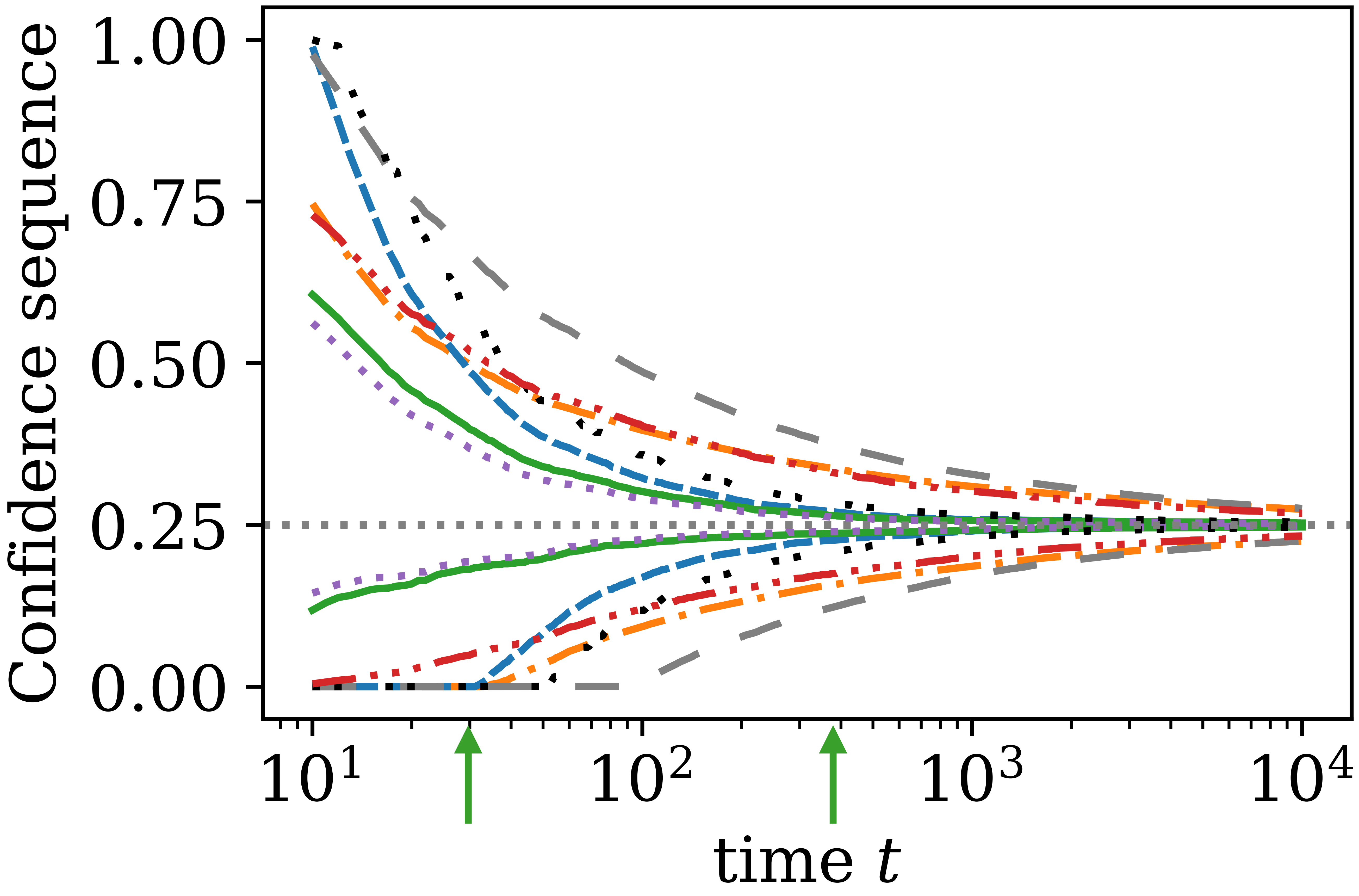
If $(C_t)_{t=1}^{\infty}$ is a confidence sequence, then C_n is a confidence interval for any fixed n.



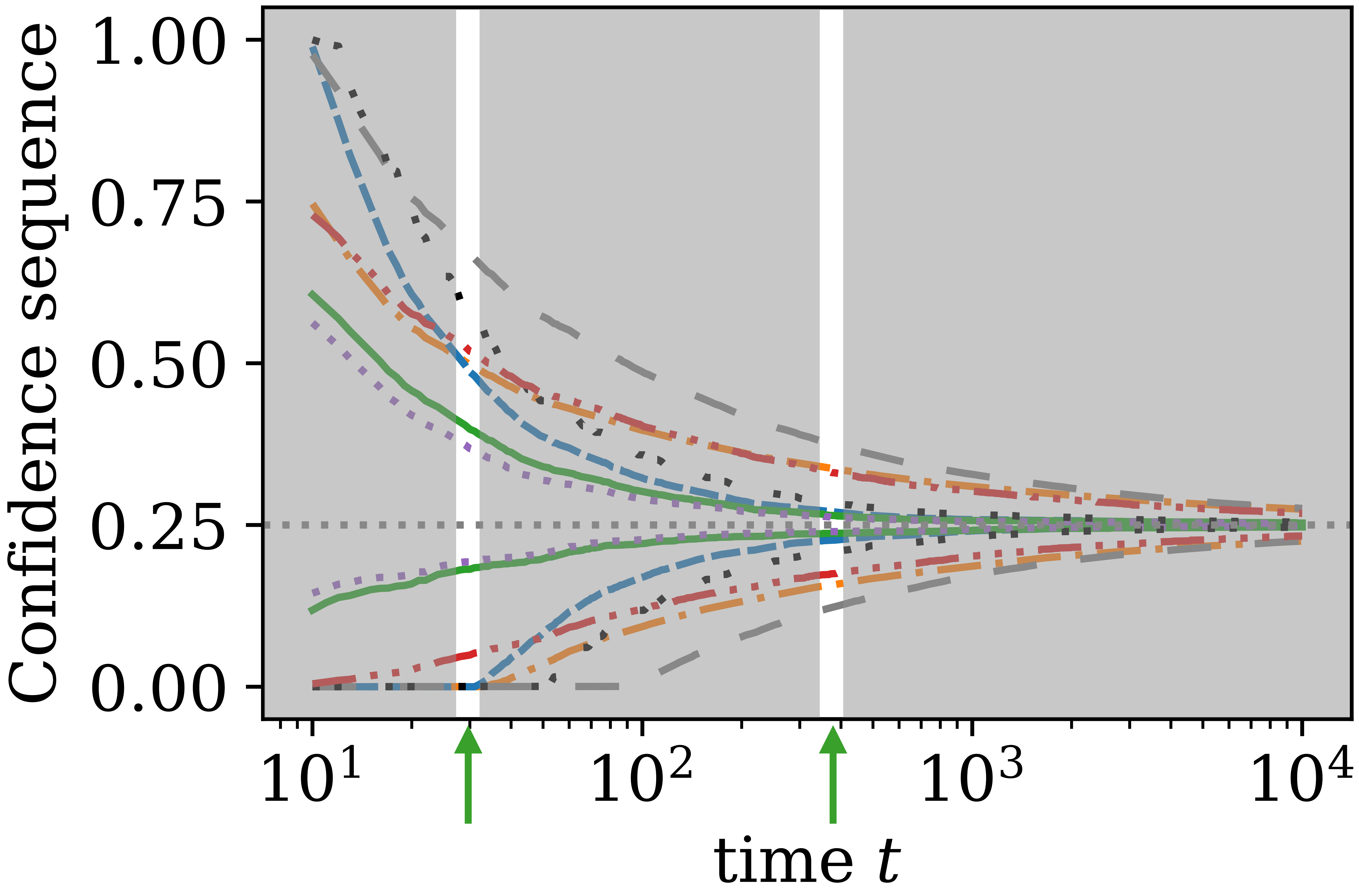
If $(C_t)_{t=1}^{\infty}$ is a confidence sequence, then C_n is a confidence interval for any fixed n.



If $(C_t)_{t=1}^{\infty}$ is a confidence sequence, then C_n is a confidence interval for any fixed n.



If $(C_t)_{t=1}^{\infty}$ is a confidence sequence, then C_n is a confidence interval for any fixed n.



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

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

After some heuristic calculations*, a good choice is

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

After some heuristic calculations*, a good choice is

$$\lambda_i^+(m) := \sqrt{\frac{2\log(2/\alpha)}{n\hat{\sigma}_{i-1}^2}} \wedge \frac{1}{m}$$

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

After some heuristic calculations*, a good choice is

$$\lambda_i^{+}(m) := \sqrt{\frac{2\log(2/\alpha)}{n\hat{\sigma}_{i-1}^2}} \wedge \frac{1}{m}$$

$$\lambda_i^{-}(m) := \sqrt{\frac{2\log(2/\alpha)}{n\hat{\sigma}_{i-1}^2}} \wedge \frac{1}{1-m}$$

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

After some heuristic calculations*, a good choice is

$$\lambda_i^{+}(m) := \sqrt{\frac{2\log(2/\alpha)}{n\hat{\sigma}_{i-1}^2}} \wedge \frac{1}{m}$$

$$\lambda_i^{-}(m) := \sqrt{\frac{2\log(2/\alpha)}{n\hat{\sigma}_{i-1}^2}} \wedge \frac{1}{1-m}$$

$$K_n^{\pm}(m) := \max\left\{ \frac{1}{2}\prod_{i=1}^{n}(1 + \lambda_i^{+} \cdot (X_i - m)), \frac{1}{2}\prod_{i=1}^{n}(1 - \lambda_i^{-} \cdot (X_i - m)) \right\}$$

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

After some heuristic calculations*, a good choice is

$$\lambda_i^+(m) := \sqrt{\frac{2\log(2/\alpha)}{n\hat{\sigma}_{i-1}^2}} \wedge \frac{1}{m}$$

$$\lambda_i^-(m) := \sqrt{\frac{2\log(2/\alpha)}{n\hat{\sigma}_{i-1}^2}} \wedge \frac{1}{1-m}$$

$$K_n^{\pm}(m) := \max\left\{ \frac{1}{2}\prod_{i=1}^{n}(1 + \lambda_i^+ \cdot (X_i - m)), \frac{1}{2}\prod_{i=1}^{n}(1 - \lambda_i^- \cdot (X_i - m)) \right\}$$

data-dependent tuning parameters without sample splitting!

So, given $X_1, \ldots, X_n$ bounded,

$$C_n^{\pm} := \left\{ m \in [0,1] : K_n^{\pm}(m) < \frac{1}{\alpha} \right\}$$

is a sharp confidence interval for μ.

So, given $X_1, \ldots, X_n$ bounded,

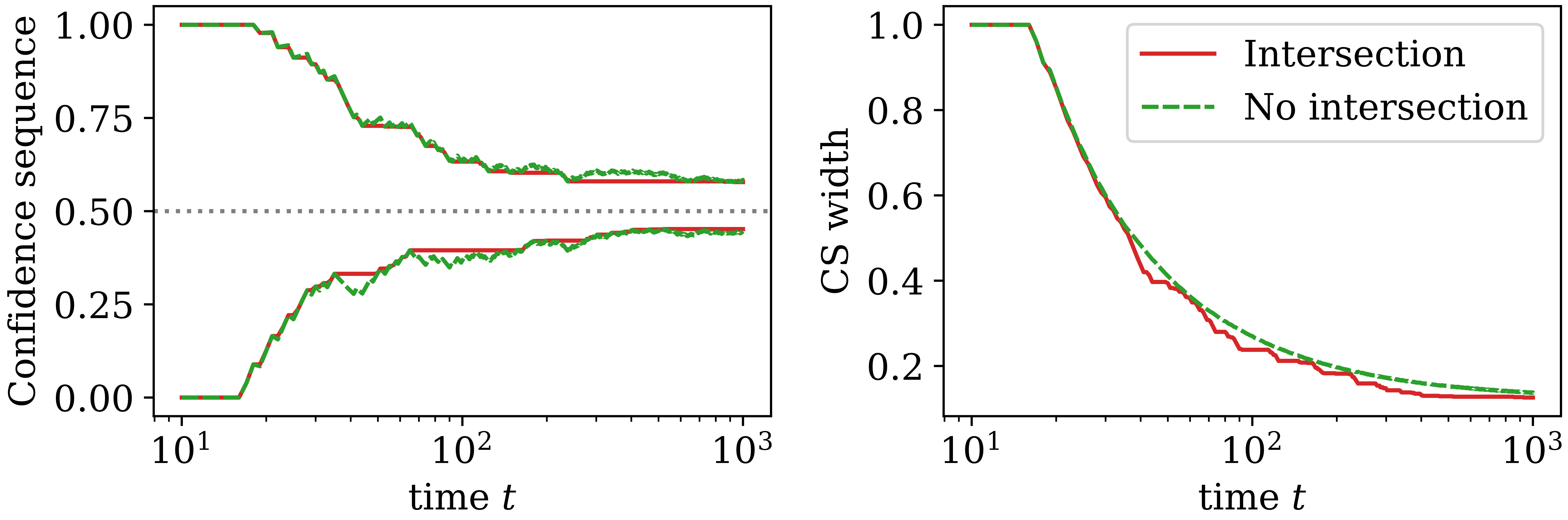
$$C_n^{\pm} := \left\{ m \in [0,1] : K_n^{\pm}(m) < \frac{1}{\alpha} \right\}$$

is a sharp confidence interval for μ.

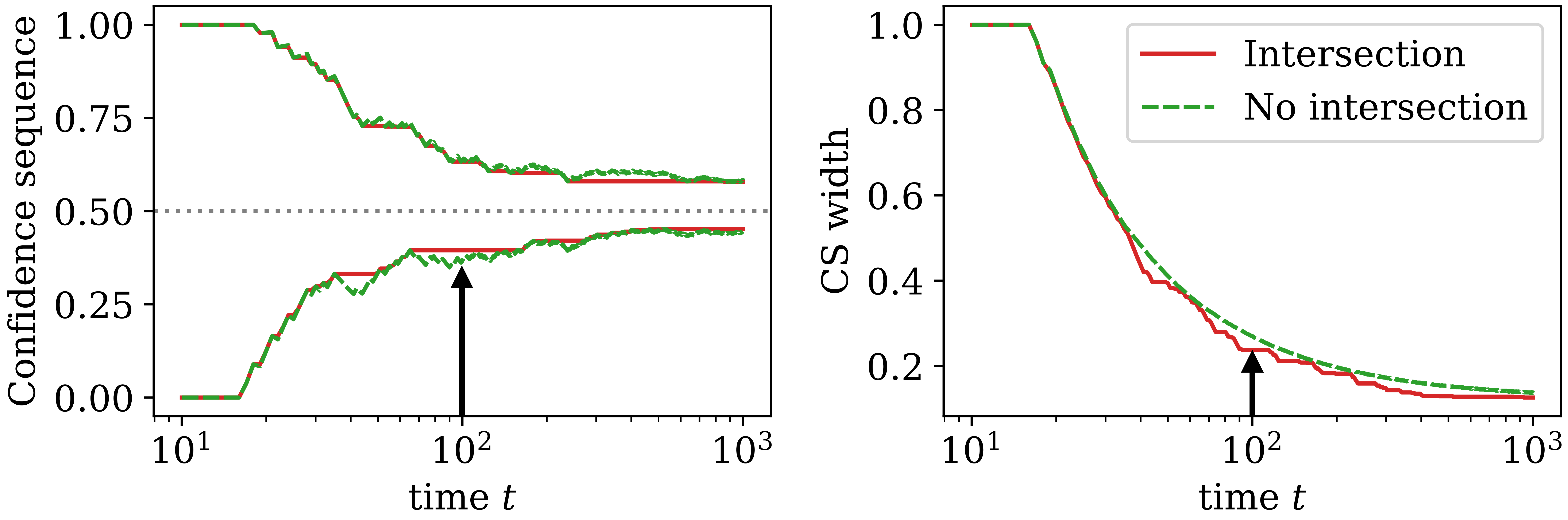
There's one more modification we can make to get **strictly** tighter confidence intervals!

If $C_1, C_2, \ldots, C_n, \ldots$ forms a $(1-\alpha)$-confidence sequence, then so does $\bigcap_{i \leq n} C_i$

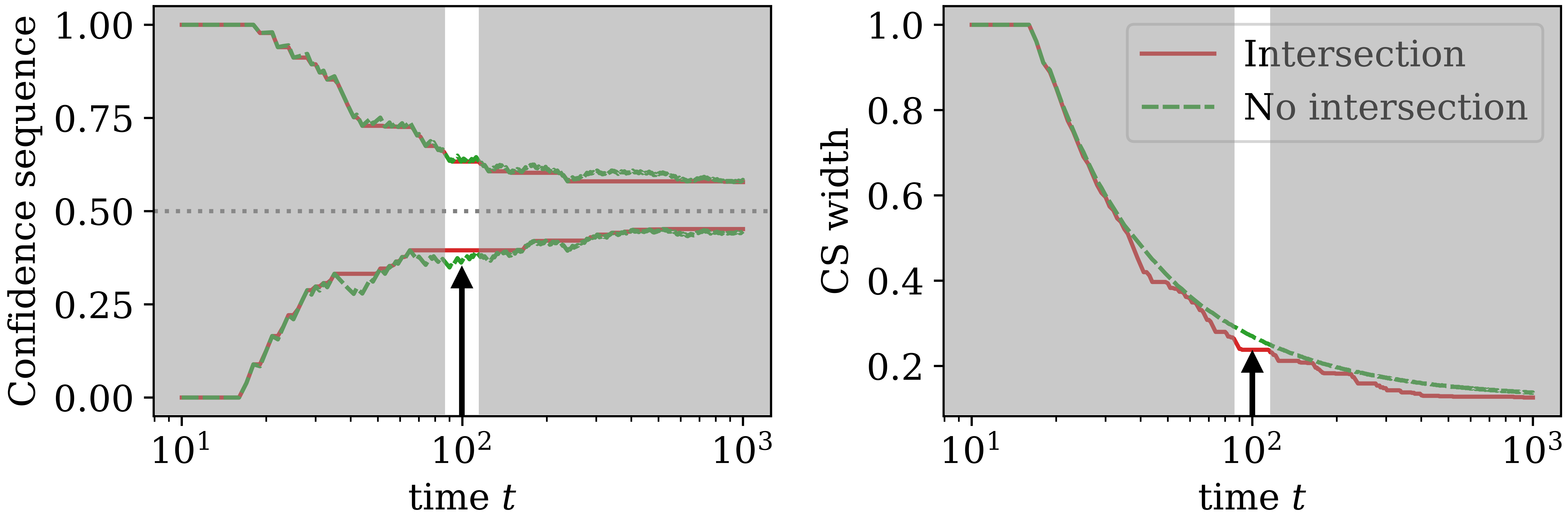
If $C_1, C_2, \ldots, C_n, \ldots$ forms a $(1-\alpha)$-confidence sequence, then so does $\bigcap_{i \le n} C_i$



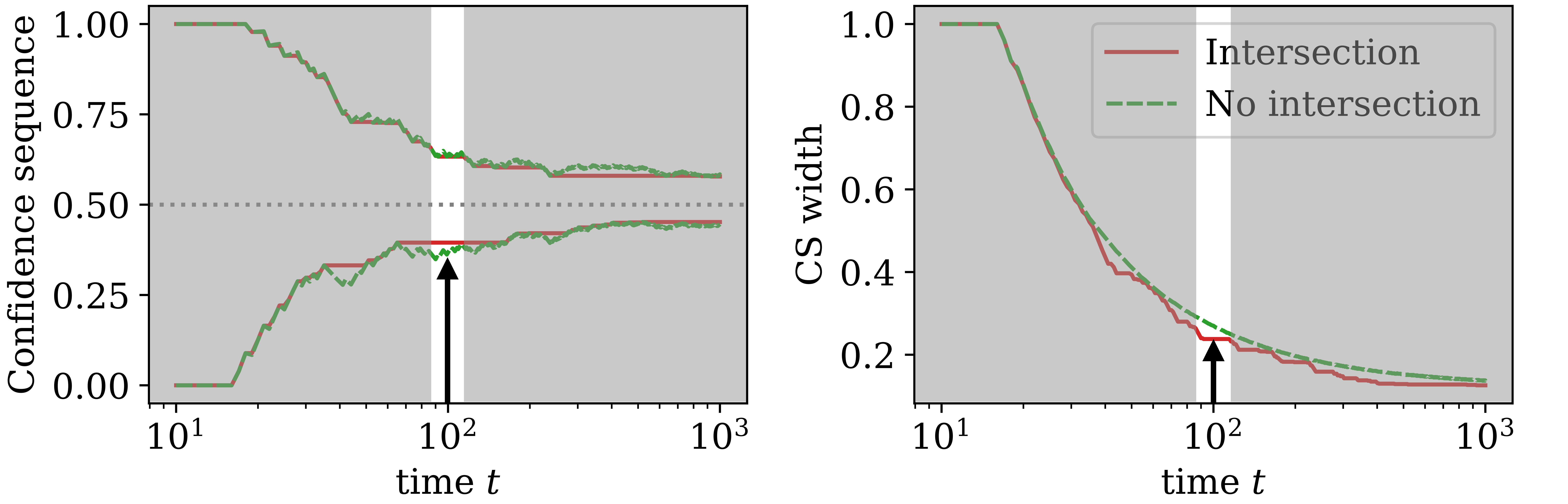
If $C_1, C_2, \ldots, C_n, \ldots$ forms a $(1-\alpha)$-confidence sequence, then so does $\bigcap_{i \leq n} C_i$



If $C_1, C_2, \ldots, C_n, \ldots$ forms a $(1-\alpha)$-confidence sequence, then so does $\displaystyle\bigcap_{i \le n} C_i$

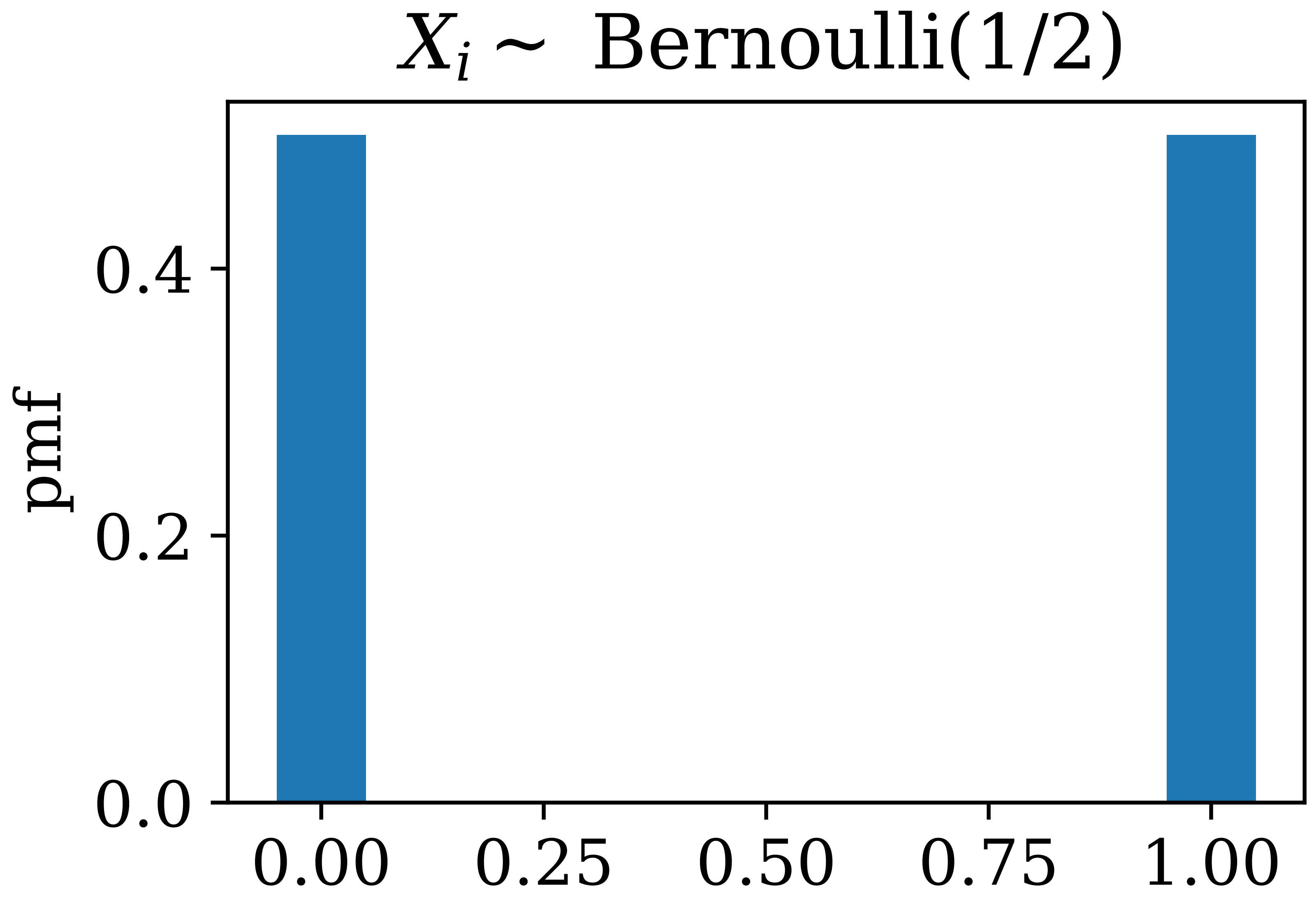


If $C_1, C_2, \ldots, C_n, \ldots$ forms a $(1-\alpha)$-confidence sequence, then so does $\bigcap_{i \le n} C_i$

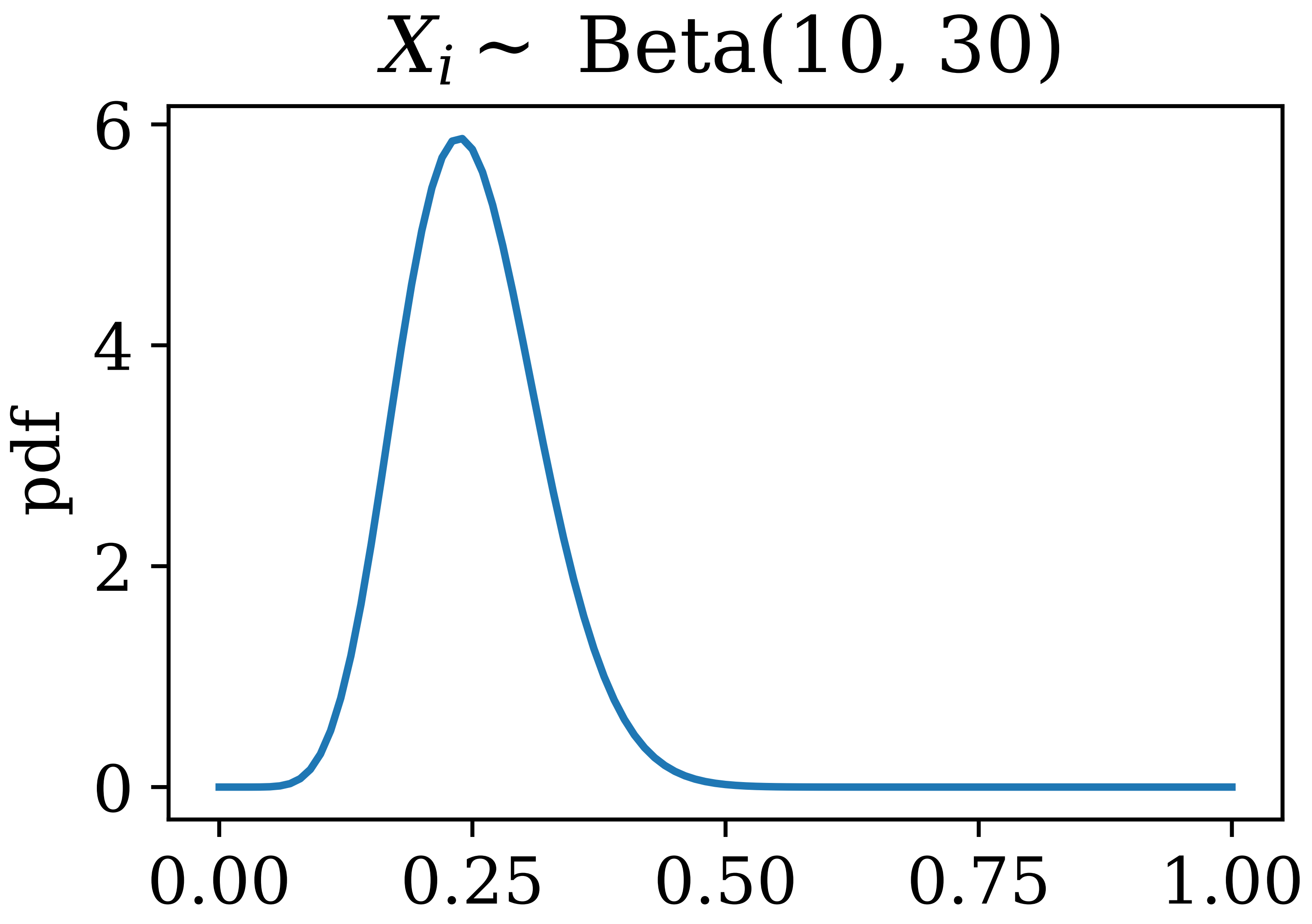


So, $\bigcap_{i \le n} C_i^{\pm}$ is a strict improvement over $C_n^{\pm}$ **for free!**

$\sigma^2 = 1/4$

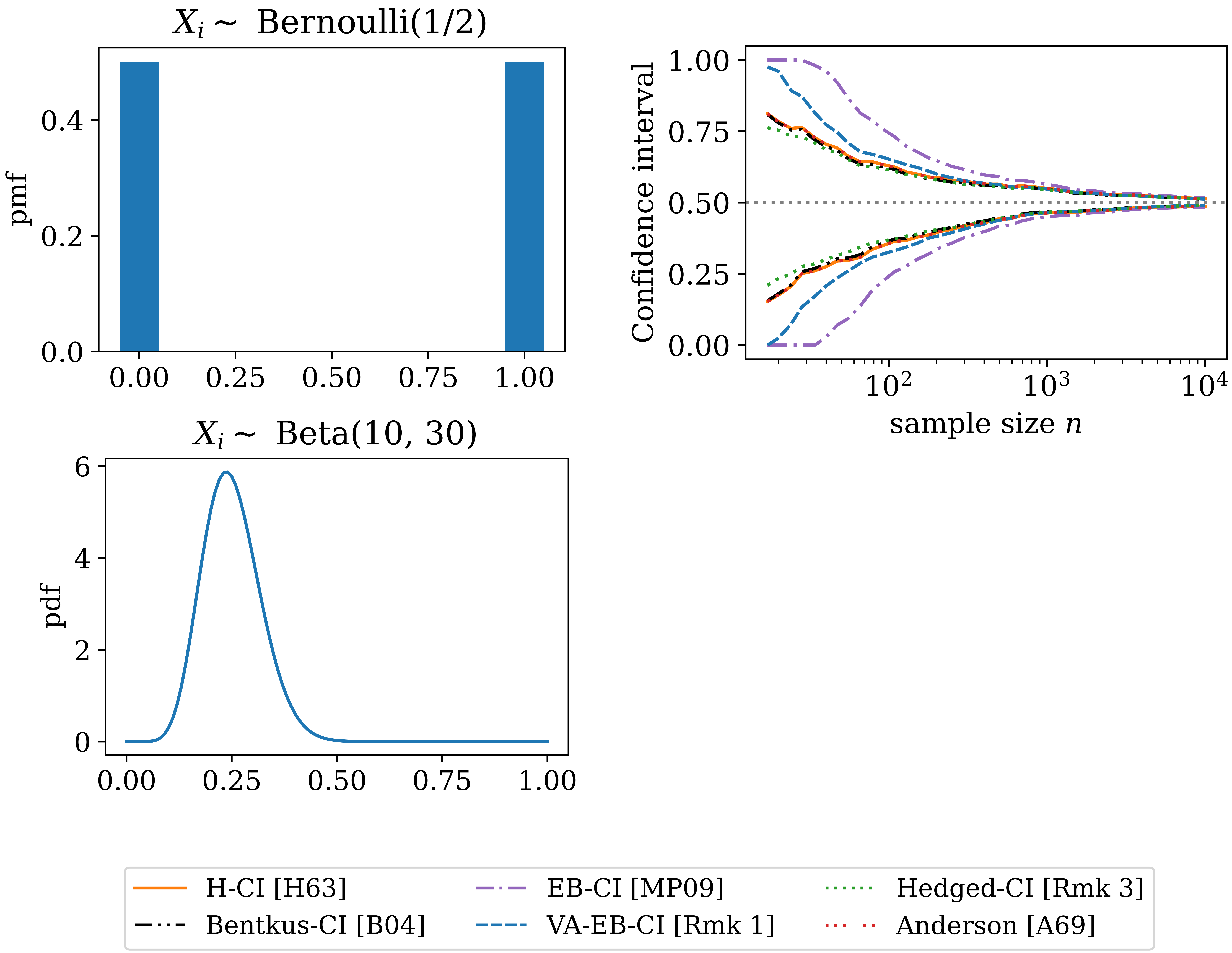


$\sigma^2 \approx 0.0046$

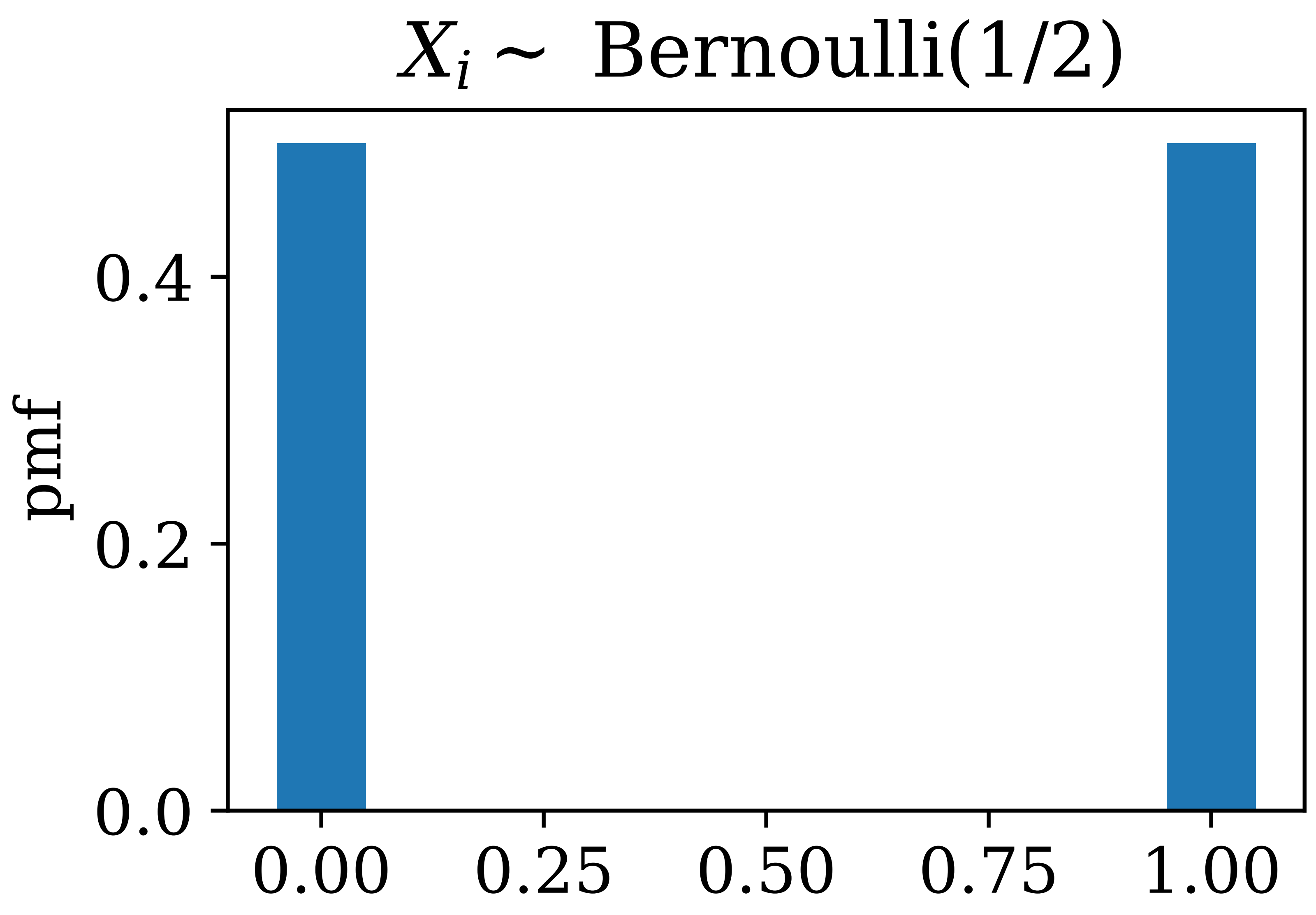


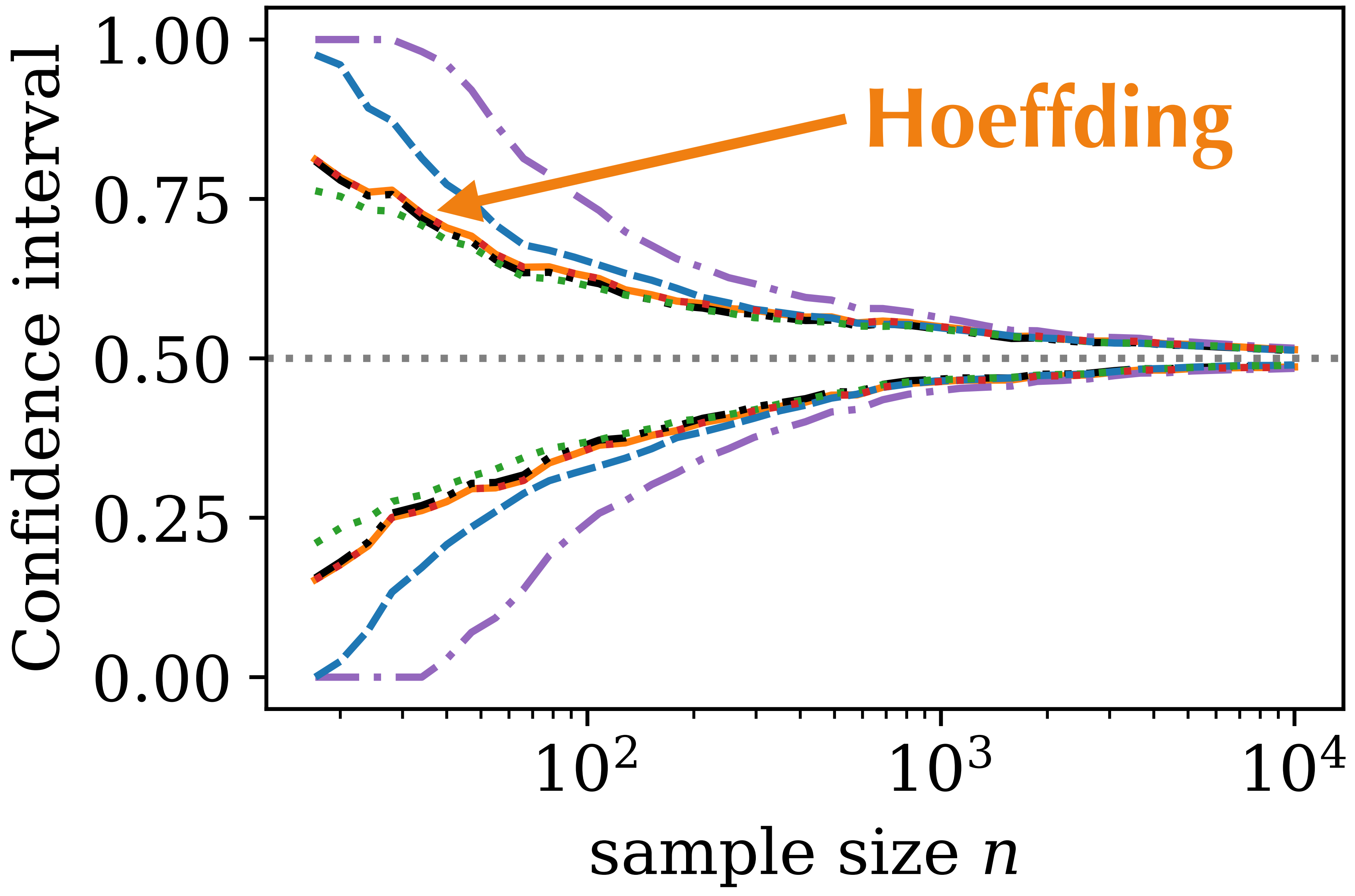
$\sigma^2 = 1/4$

$\sigma^2 \approx 0.0046$

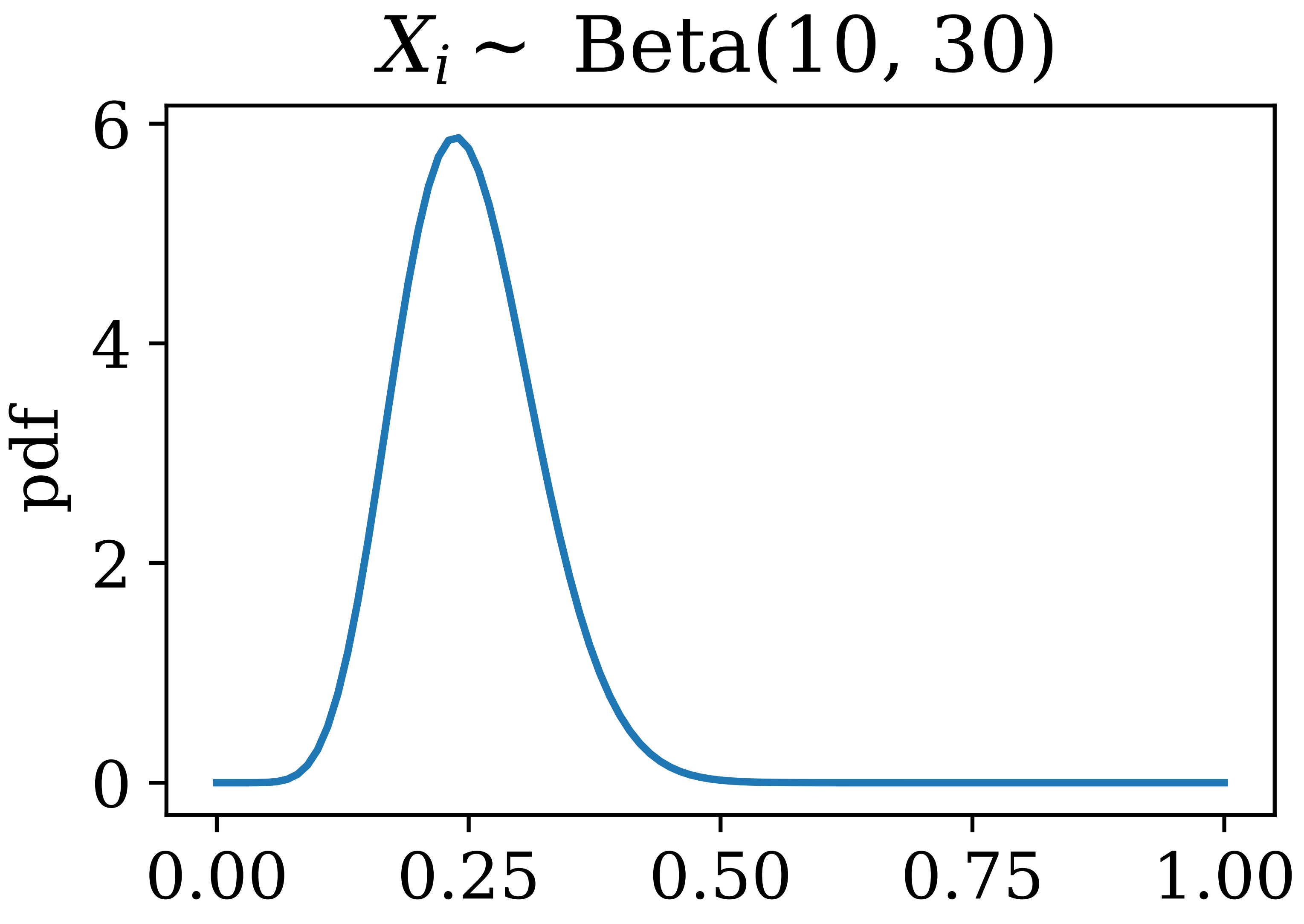


$\sigma^2 = 1/4$

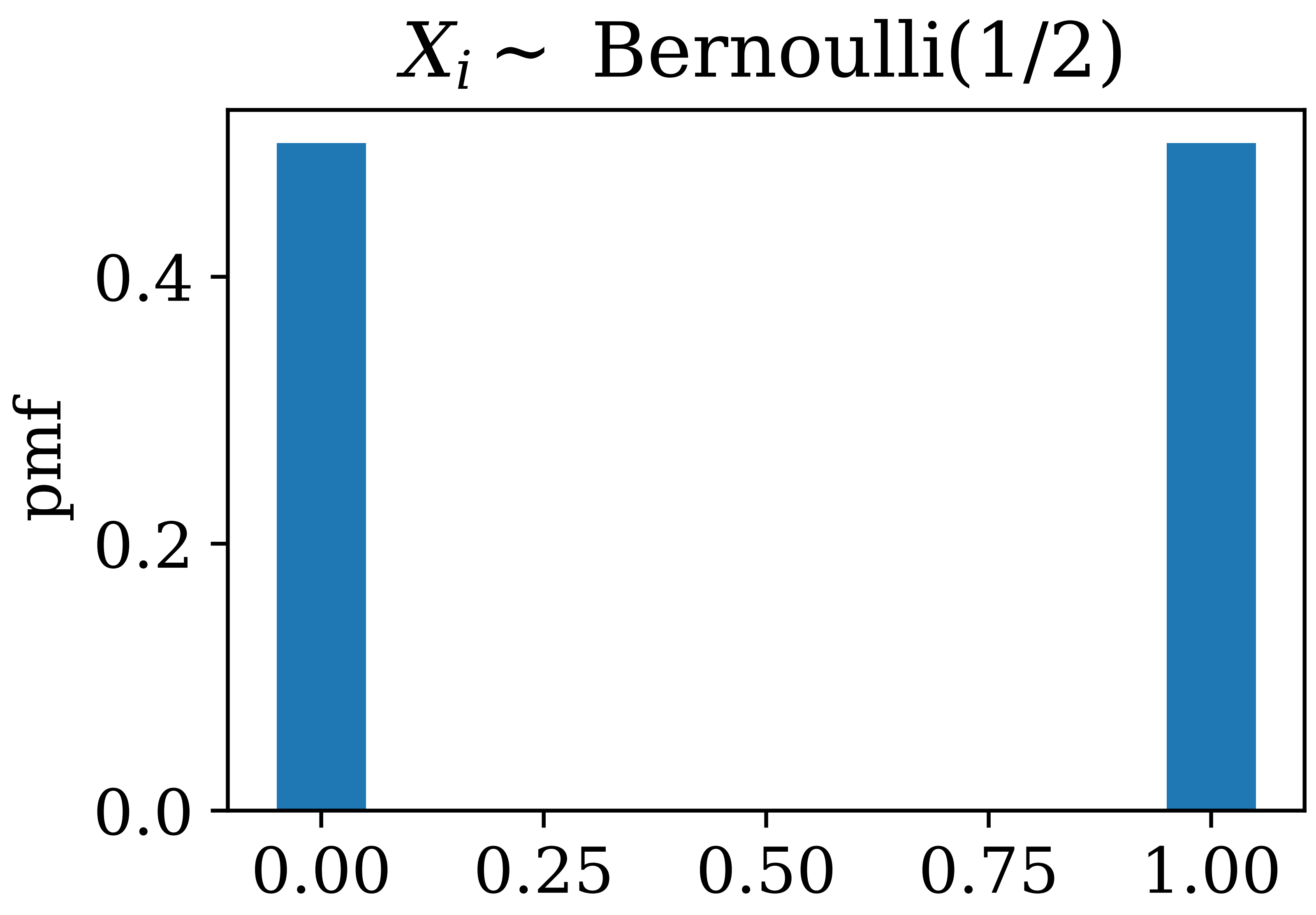


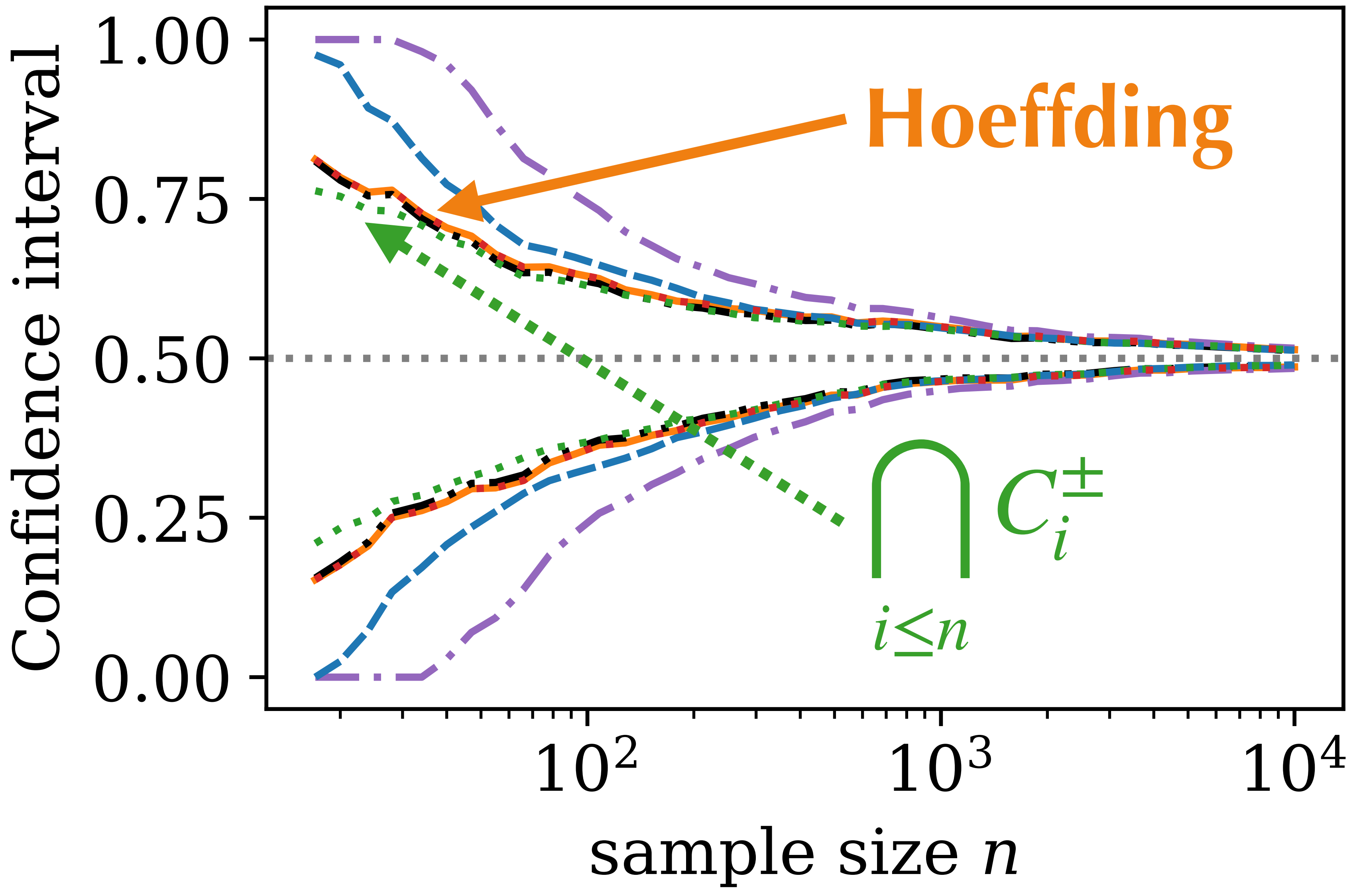


$\sigma^2 \approx 0.0046$

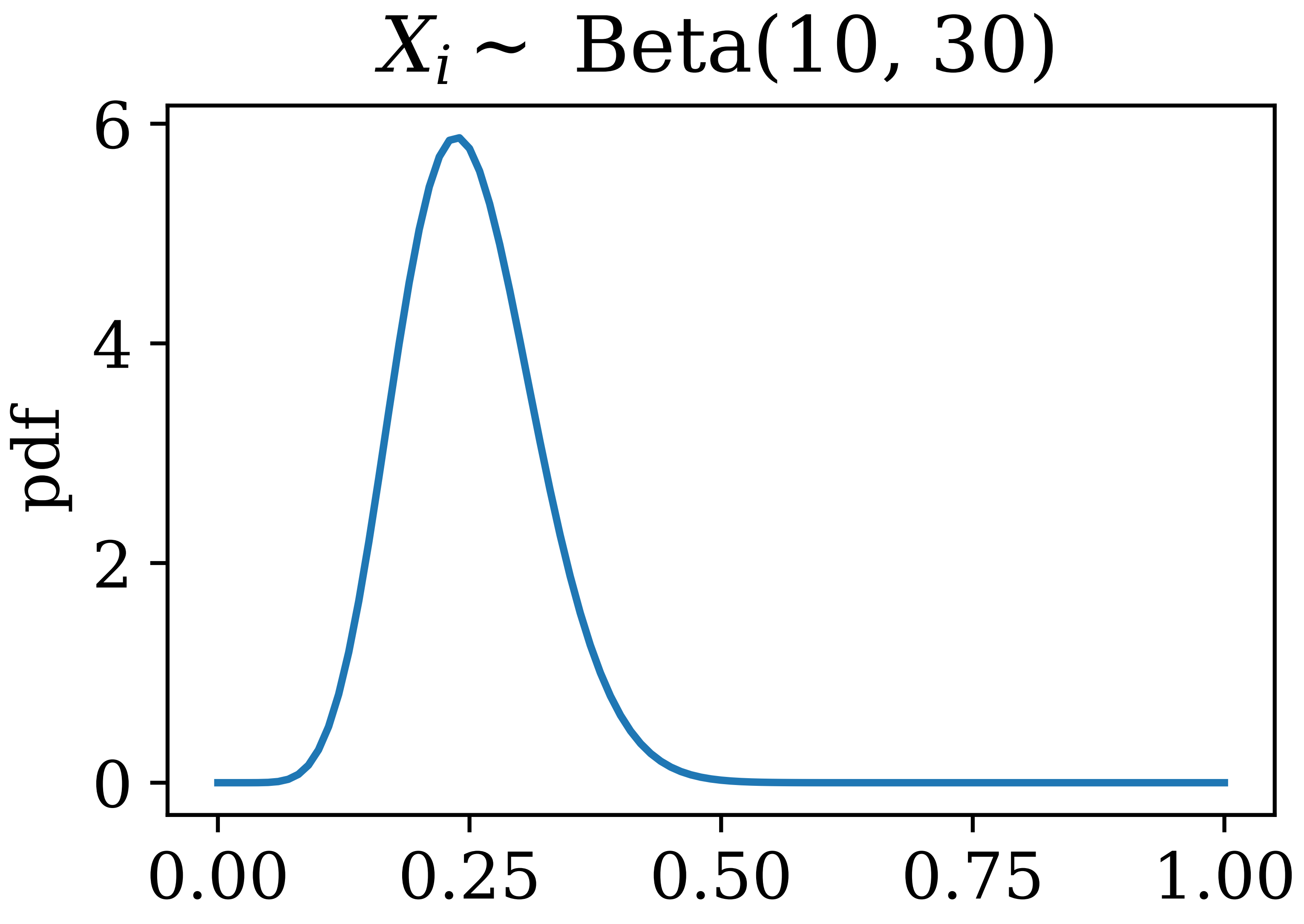


$\sigma^2 = 1/4$



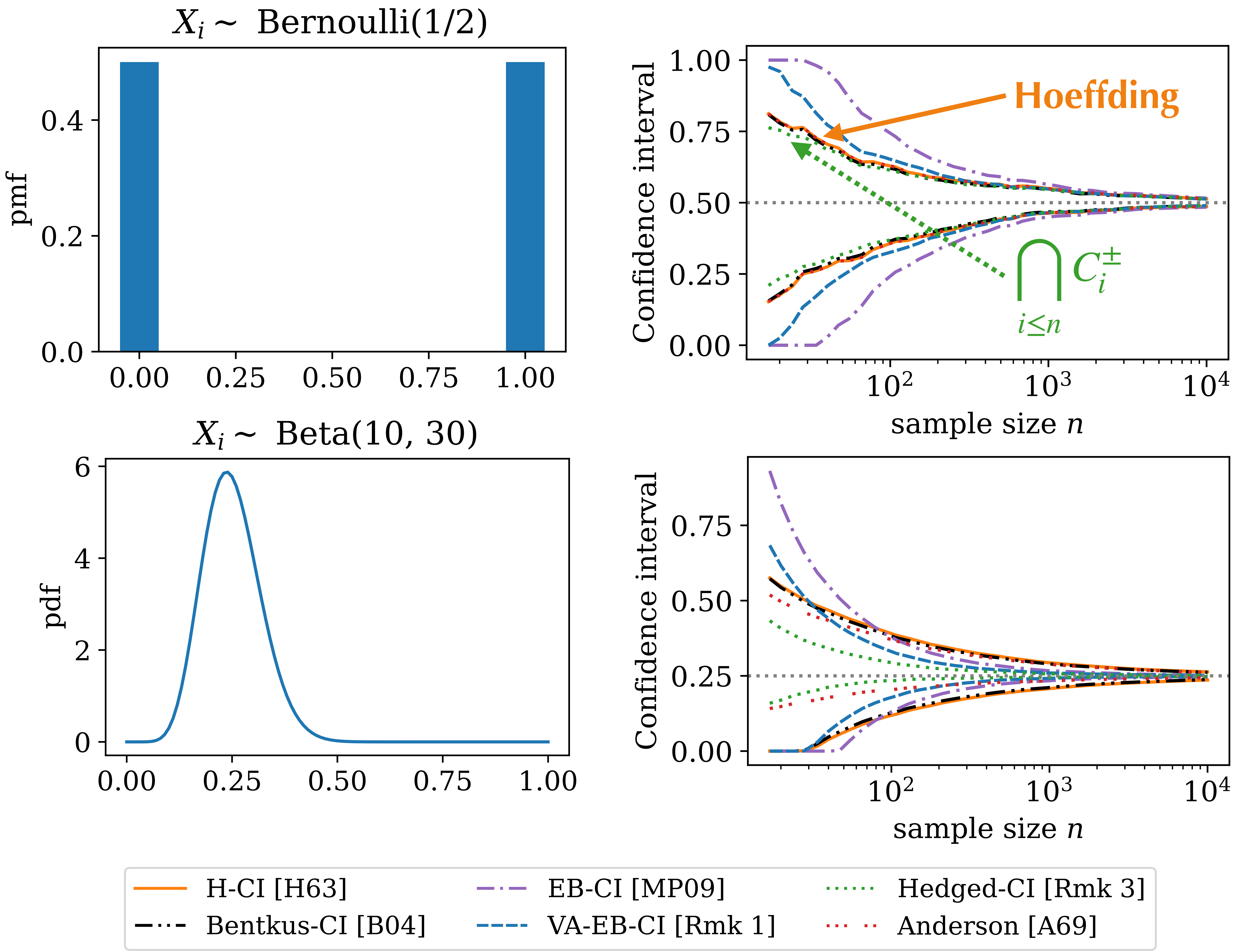


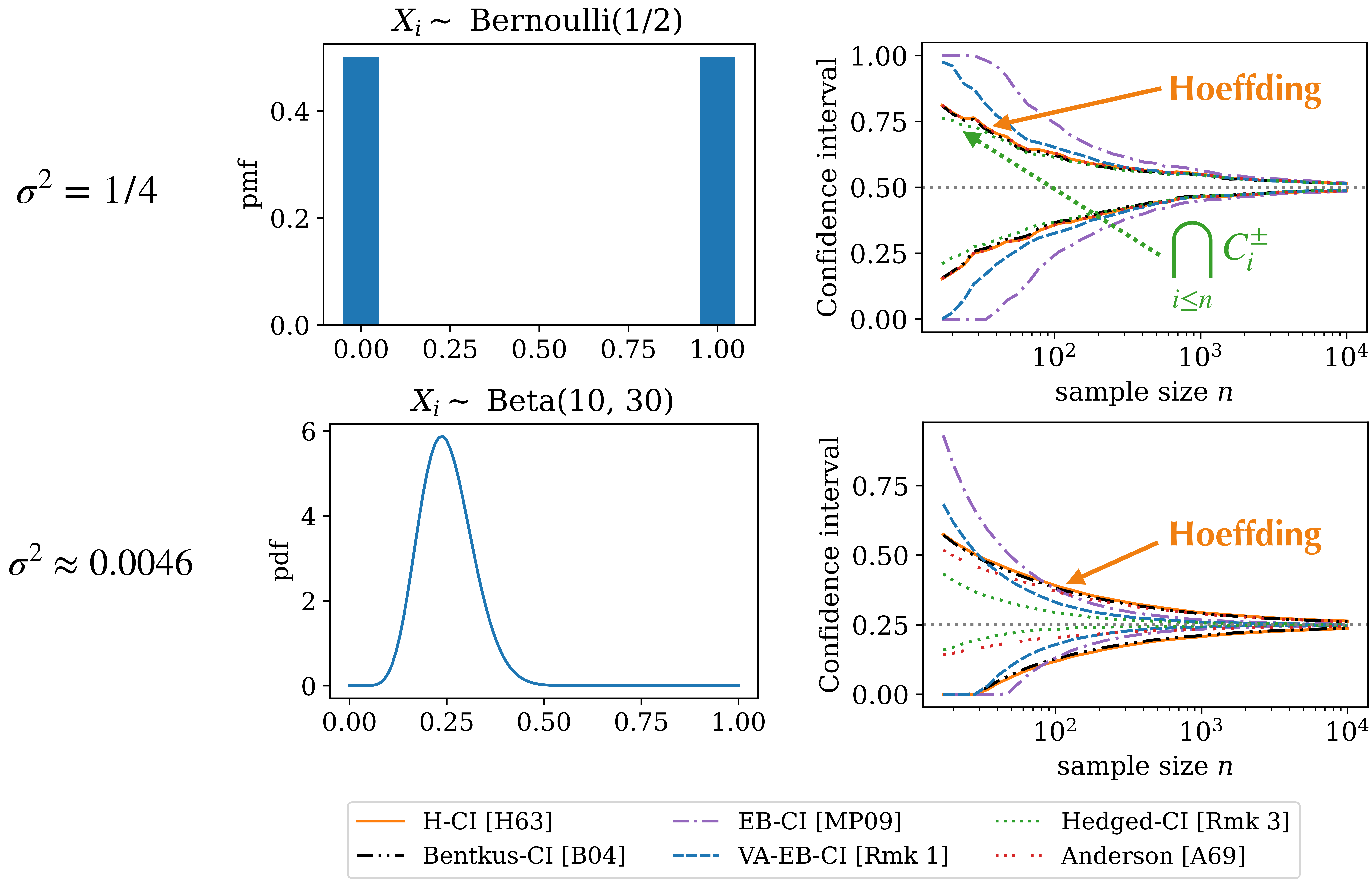
$\sigma^2 \approx 0.0046$



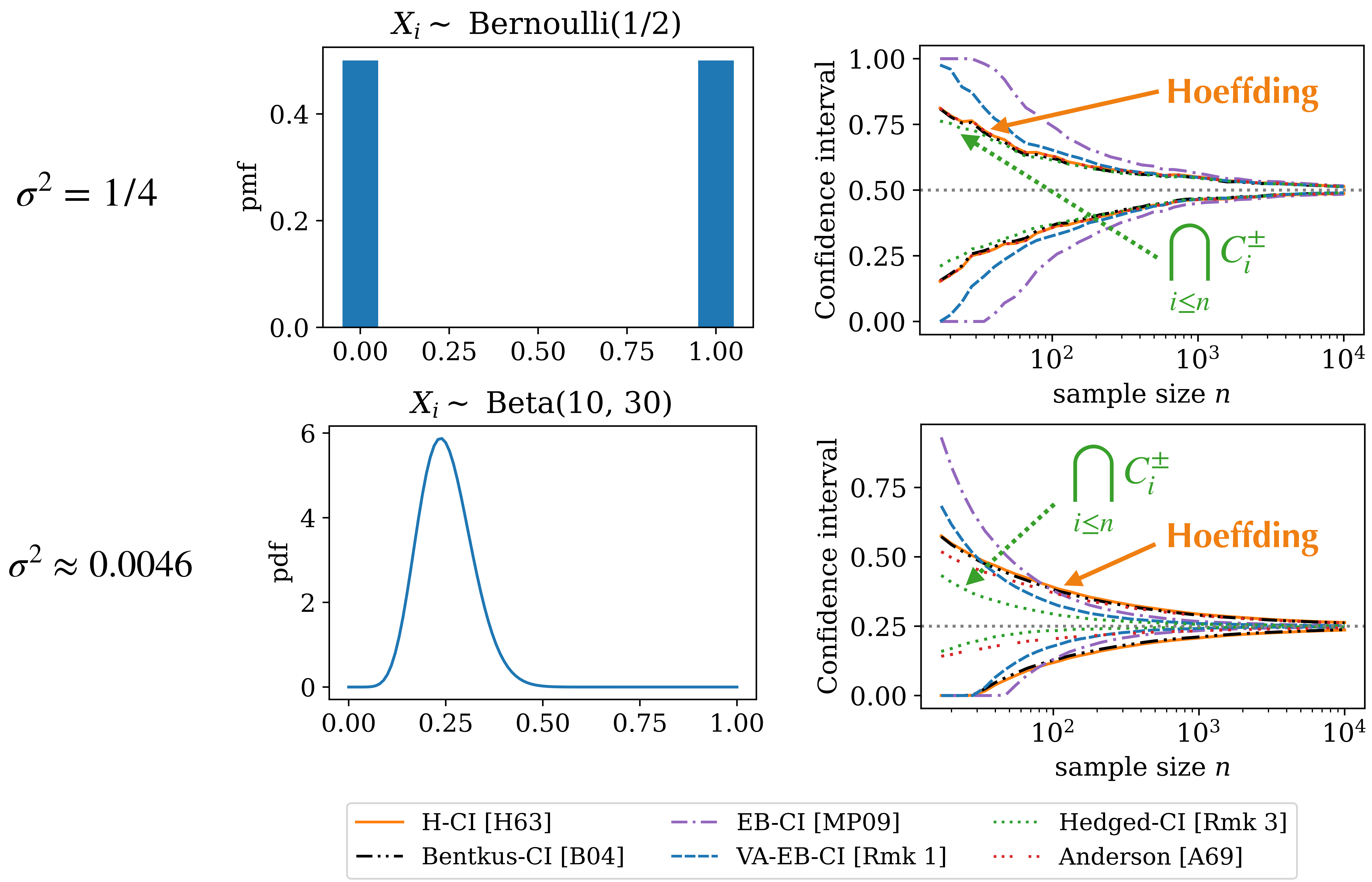
$\sigma^2 = 1/4$

$\sigma^2 \approx 0.0046$





$\sigma^2 = 1/4$
$X_i \sim$ Bernoulli(1/2)
pmf
0.0
0.2
0.4
0.00
0.25
0.50
0.75
1.00
Confidence interval
1.00
0.75
0.50
0.25
0.00
Hoeffding
$\bigcap_{i \le n} C_i^{\pm}$
10^2
10^3
10^4
sample size n
$\sigma^2 \approx 0.0046$
$X_i \sim$ Beta(10, 30)
pdf
0
2
4
6
0.00
0.25
0.50
0.75
1.00
Confidence interval
0.75
0.50
0.25
0.00
Hoeffding
10^2
10^3
10^4
sample size n
H-CI [H63]
Bentkus-CI [B04]
EB-CI [MP09]
VA-EB-CI [Rmk 1]
Hedged-CI [Rmk 3]
Anderson [A69]



$\sigma^2 = 1/4$
$X_i \sim$ Bernoulli(1/2)
pmf
Confidence interval
Hoeffding
$\bigcap_{i \le n} C_i^\pm$
sample size n
$\sigma^2 \approx 0.0046$
$X_i \sim$ Beta(10, 30)
pdf
Confidence interval
$\bigcap_{i \le n} C_i^\pm$
Hoeffding
sample size n
H-CI [H63]
Bentkus-CI [B04]
EB-CI [MP09]
VA-EB-CI [Rmk 1]
Hedged-CI [Rmk 3]
Anderson [A69]

Recap of fixed-time confidence intervals:

Recap of fixed-time confidence intervals:

Used ***sequential*** ideas for ***fixed-time*** inference.
Most notably,

Recap of fixed-time confidence intervals:

Used ***sequential*** ideas for ***fixed-time*** inference.
Most notably,

1. Data-dependent tuning parameters, and

Recap of fixed-time confidence intervals:

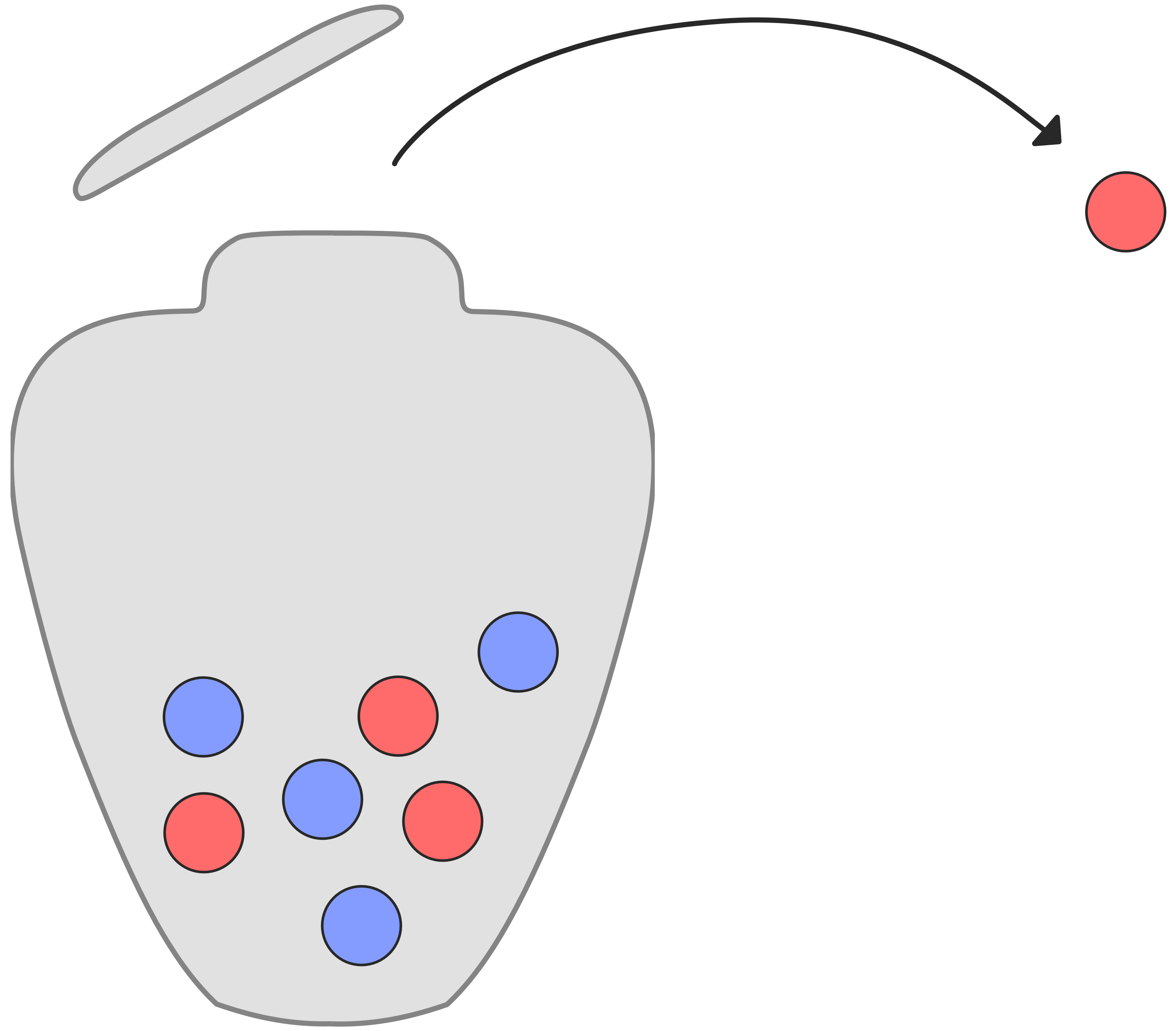
Used ***sequential*** ideas for ***fixed-time*** inference.
Most notably,

1. Data-dependent tuning parameters, and
2. A strictly-stronger-than-Markov inequality.

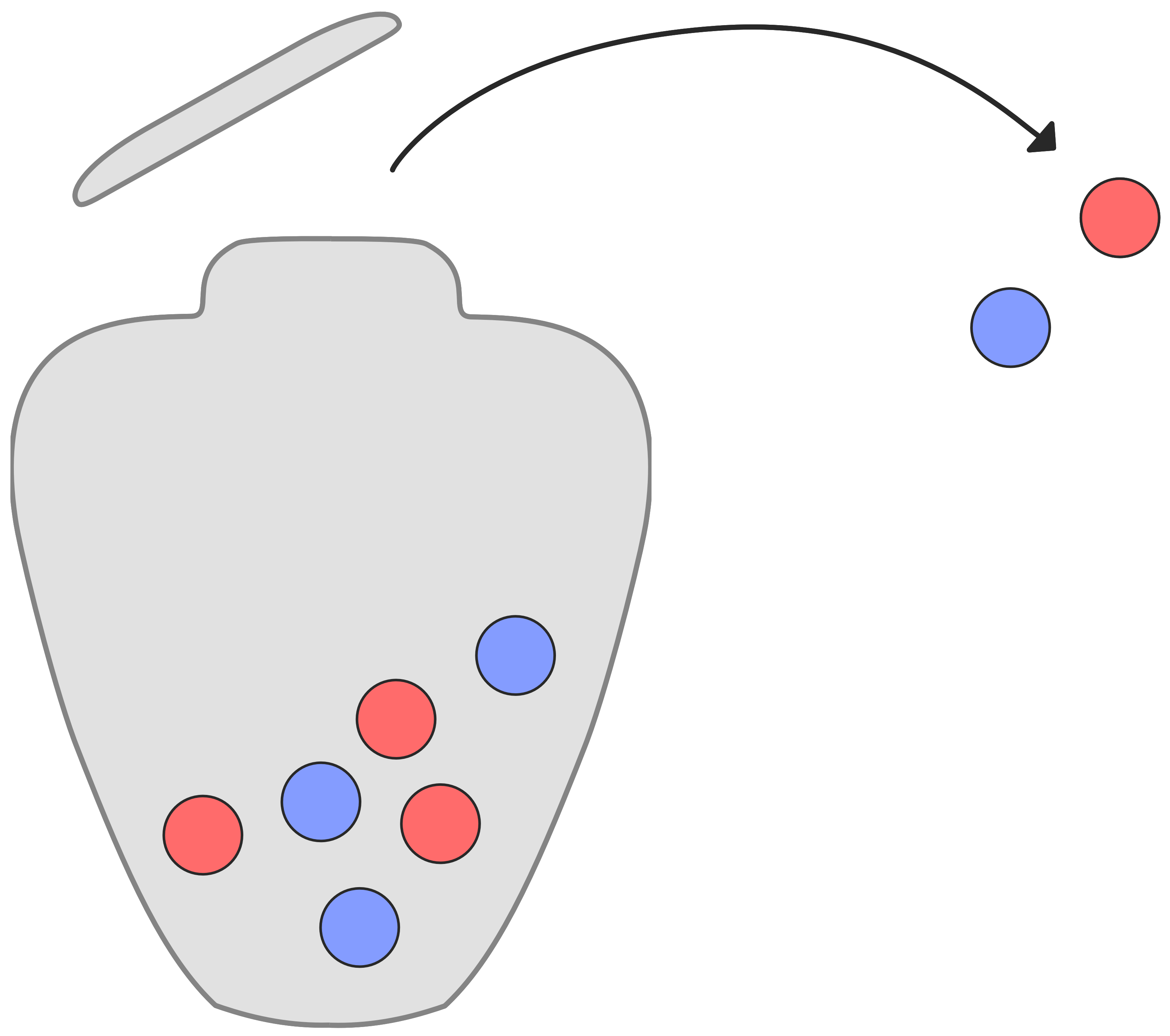
Confidence sets for sampling without replacement

Without-replacement (WoR) sampling:

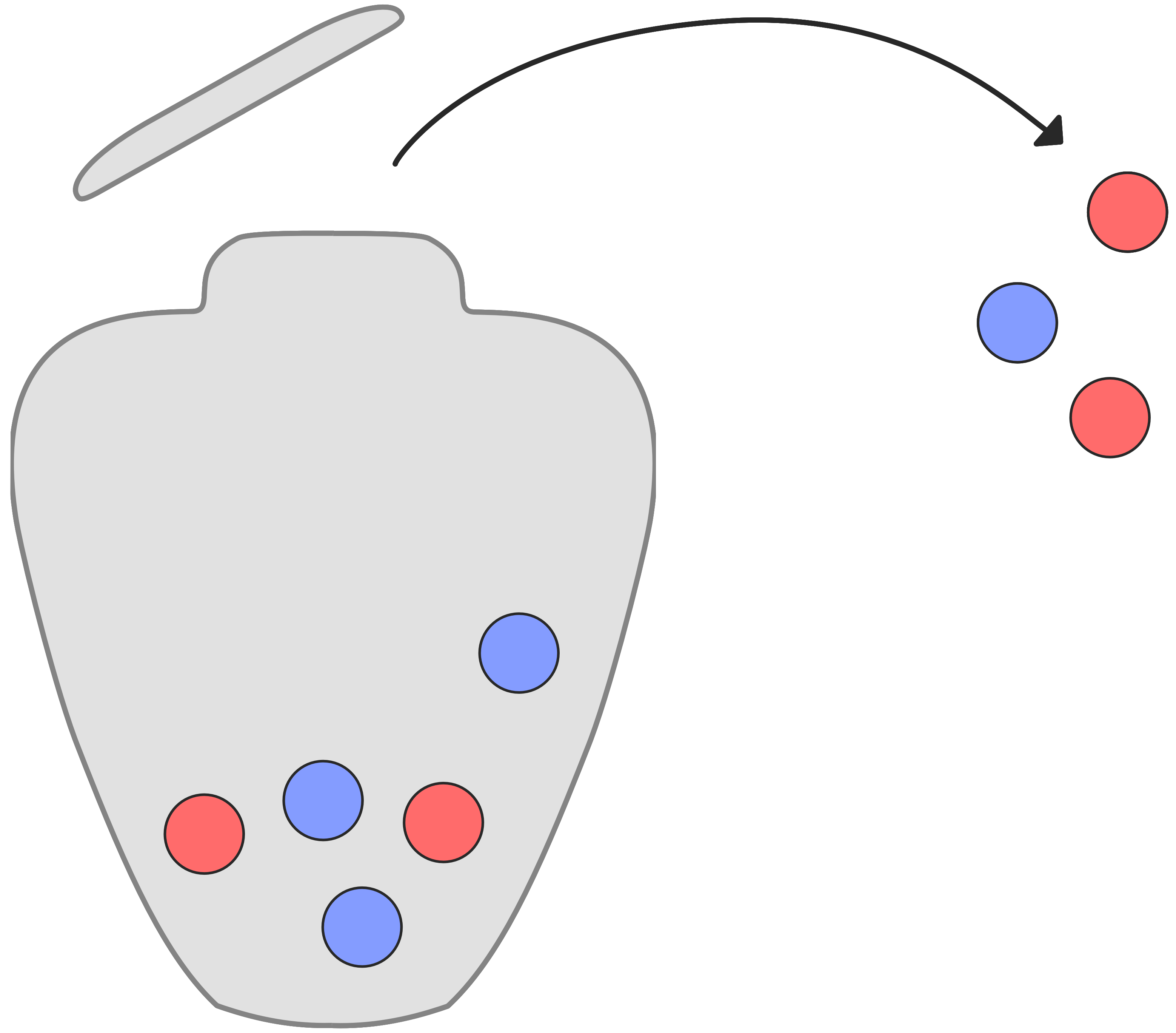
Without-replacement (WoR) sampling:



Without-replacement (WoR) sampling:

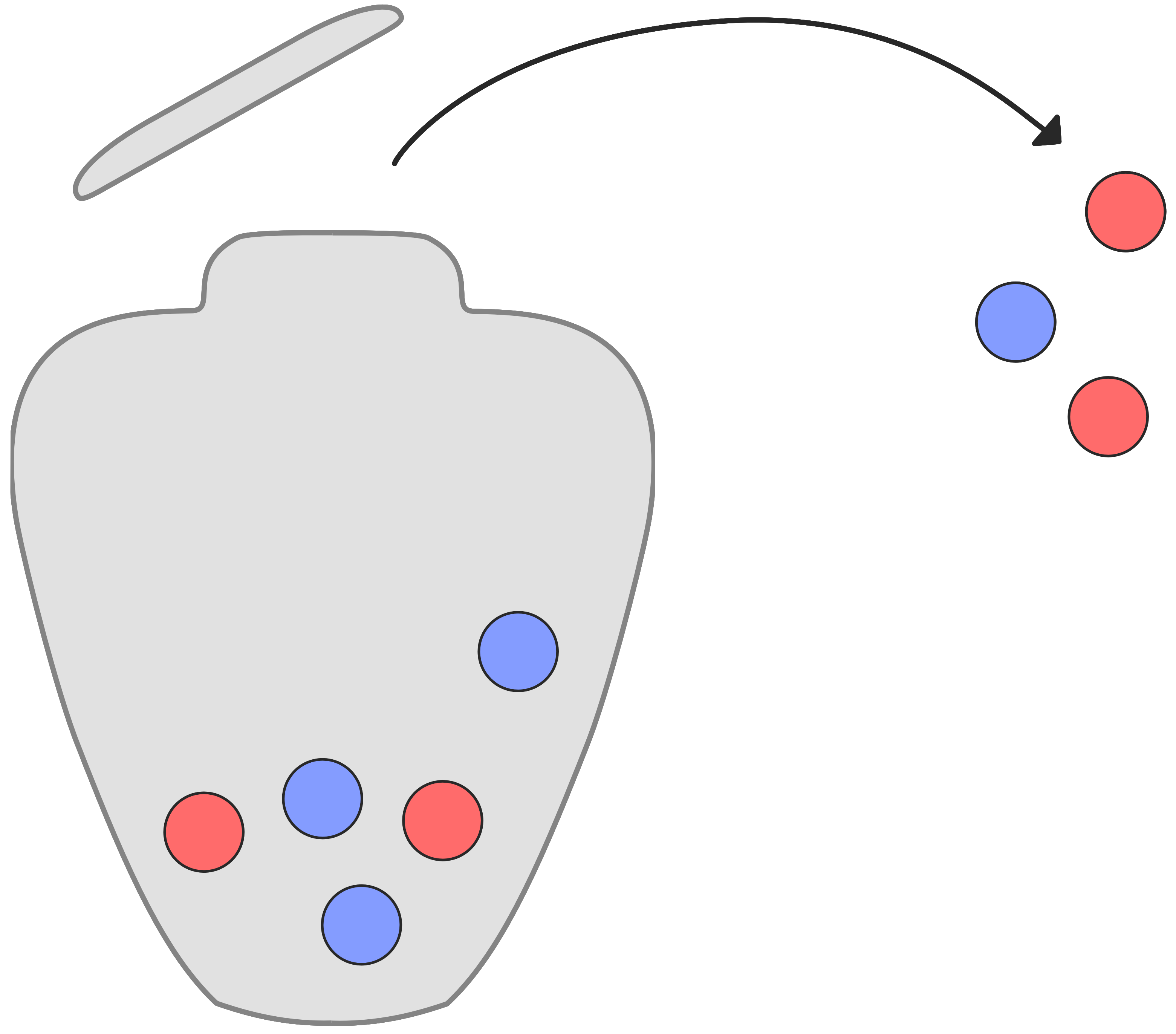


Without-replacement (WoR) sampling:

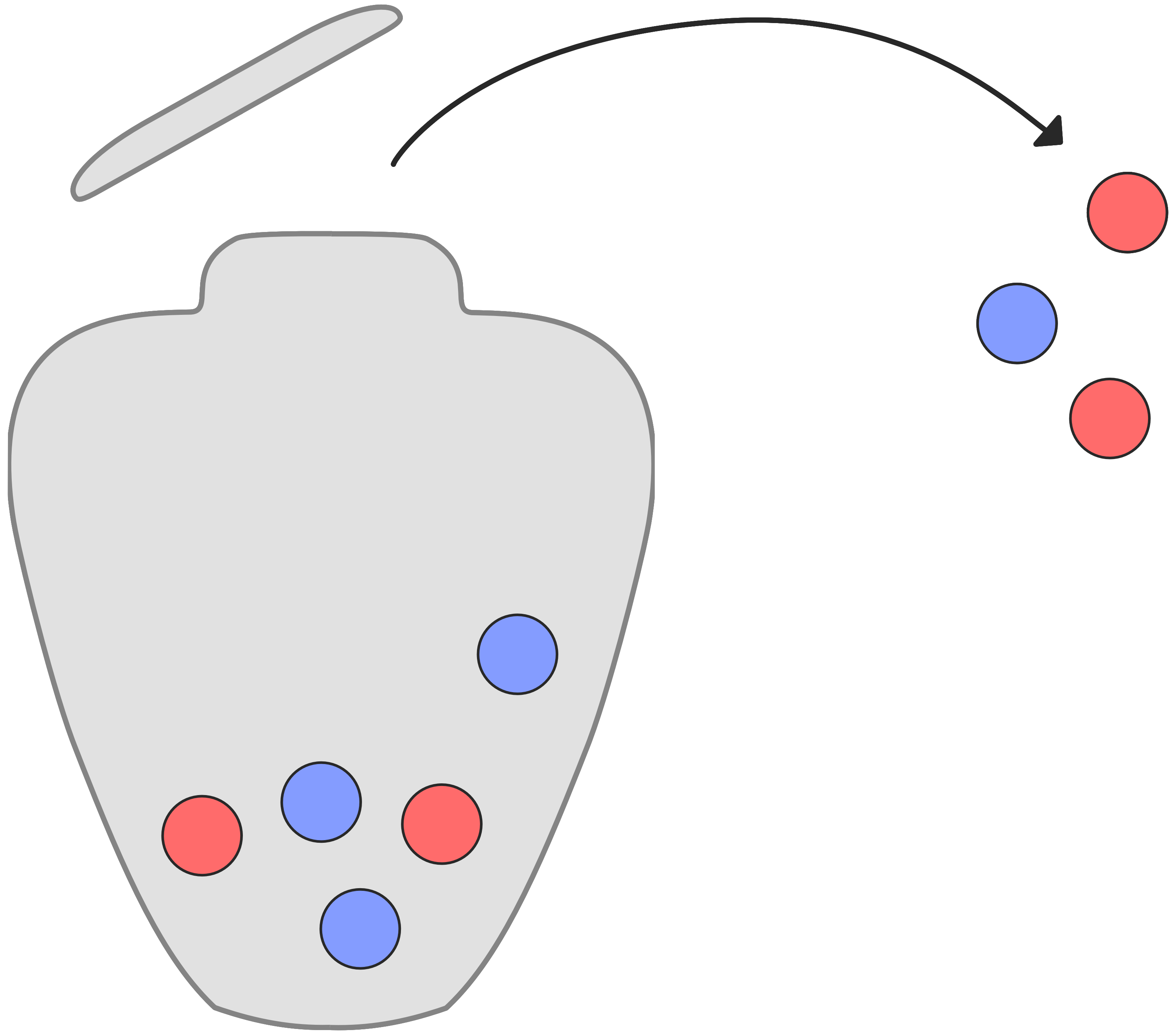


Without-replacement (WoR) sampling:

$$(x_1, \ldots, x_N) \in [0,1]^N, \ \ \mu := \frac{1}{N}\sum_{i=1}^{N} x_i$$



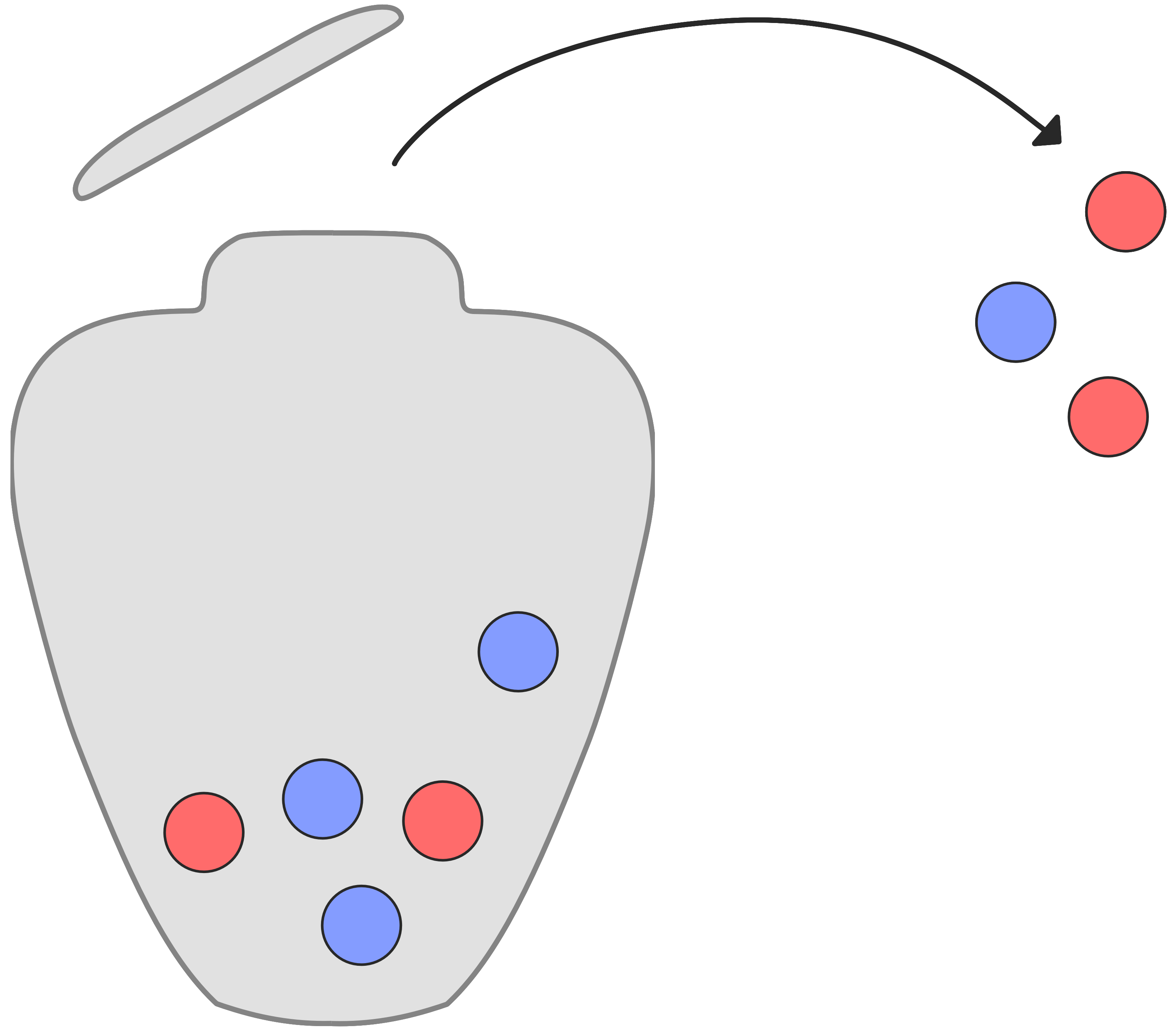
Without-replacement (WoR) sampling:



$$(x_1, \ldots, x_N) \in [0,1]^N, \;\; \mu := \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$X_1 \sim \mathrm{Unif}\left((x_1, \ldots, x_N)\right)$$

Without-replacement (WoR) sampling:

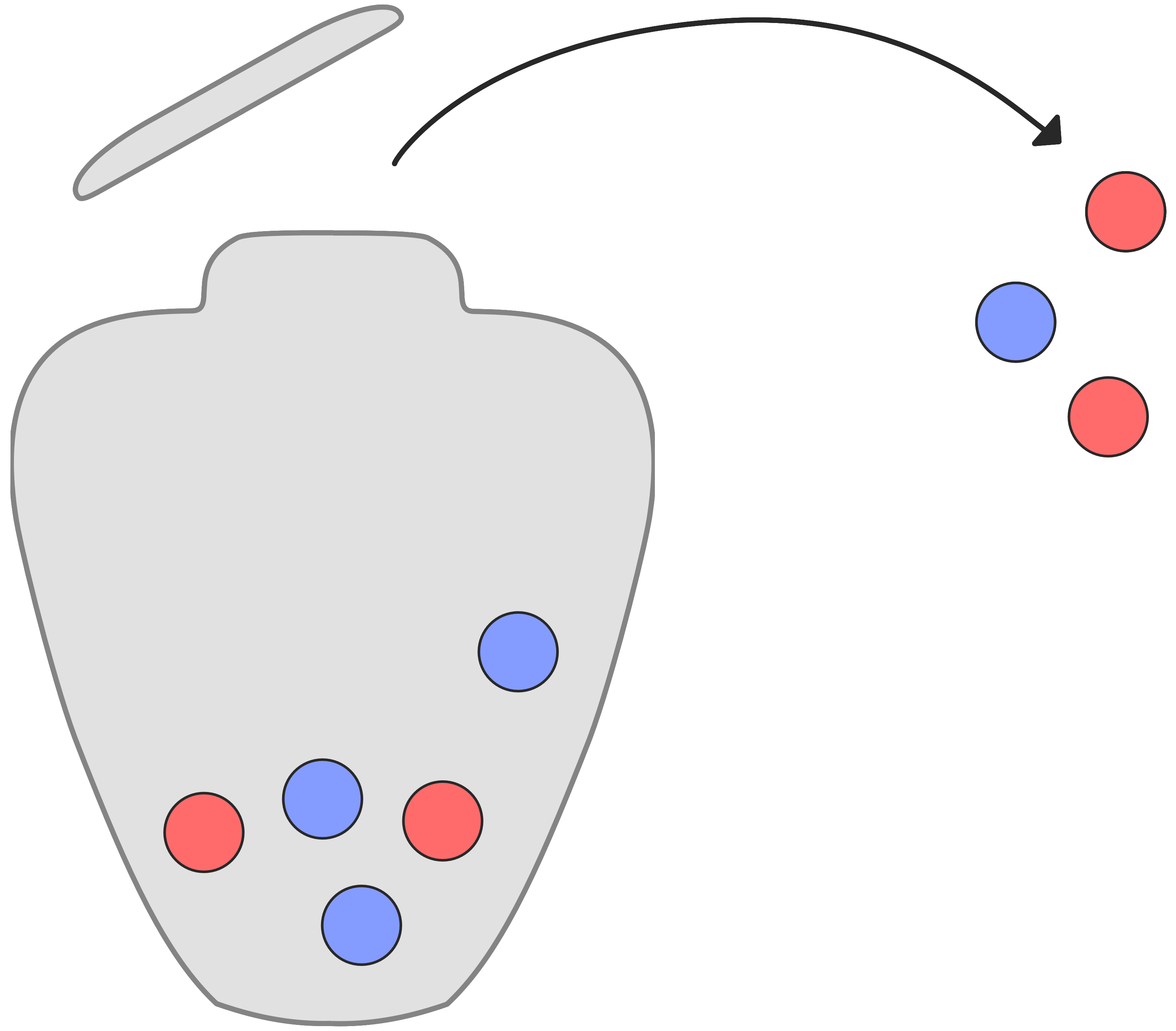


$$(x_1, \ldots, x_N) \in [0,1]^N, \quad \mu := \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$X_1 \sim \mathrm{Unif}\big((x_1, \ldots, x_N)\big)$$

$$X_2 \sim \mathrm{Unif}\big((x_1, \ldots, x_N) \backslash X_1\big)$$

Without-replacement (WoR) sampling:



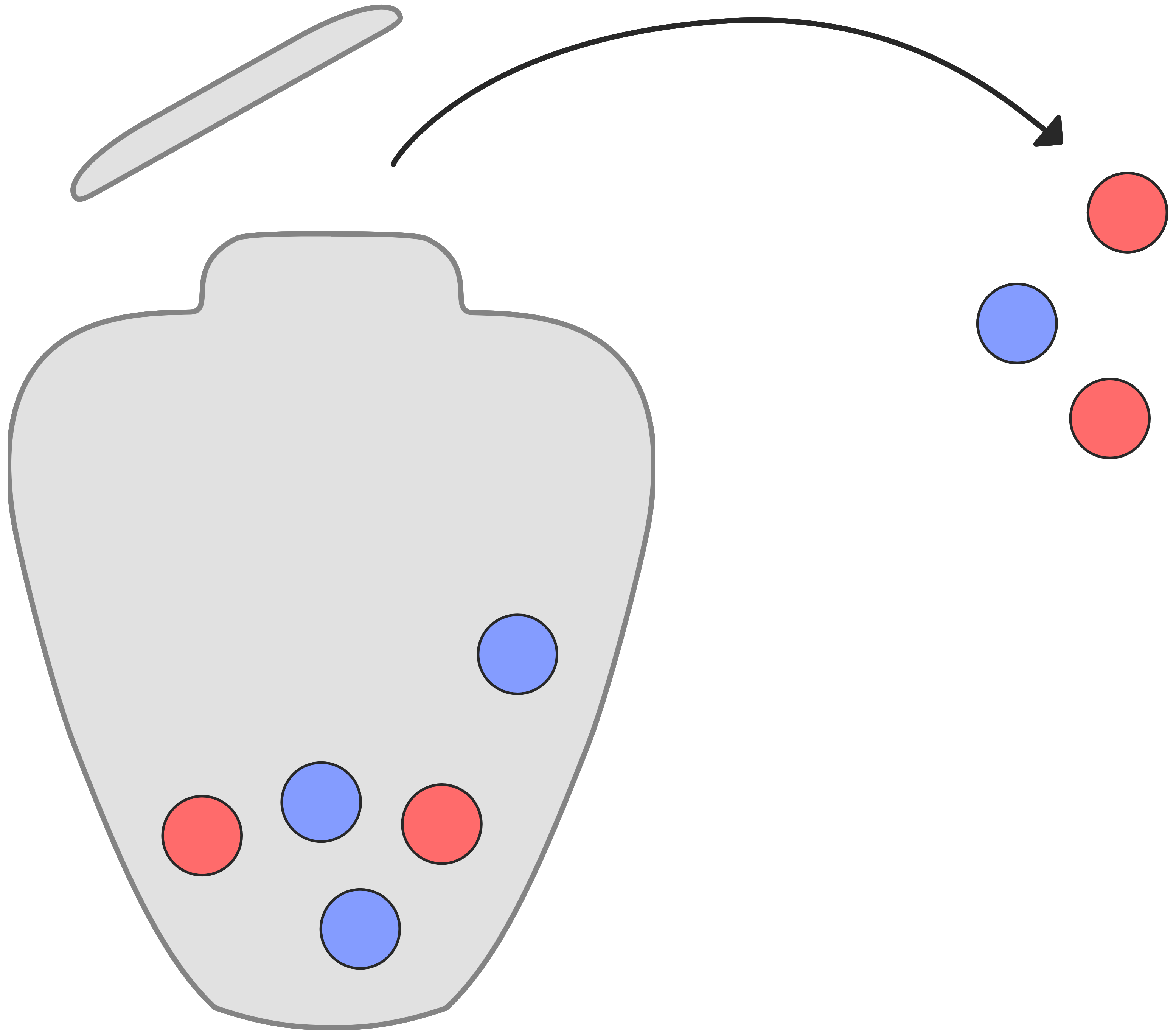
$$(x_1, \ldots, x_N) \in [0,1]^N, \;\; \mu := \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$X_1 \sim \text{Unif}\left((x_1, \ldots, x_N)\right)$$

$$X_2 \sim \text{Unif}\left((x_1, \ldots, x_N) \backslash X_1\right)$$

$$\vdots$$

Without-replacement (WoR) sampling:



$$(x_1, \ldots, x_N) \in [0,1]^N, \quad \mu := \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$X_1 \sim \text{Unif}\left((x_1, \ldots, x_N)\right)$$

$$X_2 \sim \text{Unif}\left((x_1, \ldots, x_N) \backslash X_1\right)$$

$$\vdots$$

$$X_t \sim \text{Unif}\left((x_1, \ldots, x_N) \backslash X_1^{t-1}\right)$$

Want to estimate $\mu := \frac{1}{N}\sum_{i=1}^{N} x_i$

Goal: construct a game so that $\left(K_t(\mu)\right)_{t=0}^{N}$ is a martingale under WoR sampling.

$$X_t \sim \mathrm{Unif}\left((x_1, \ldots, x_N) \backslash X_1^{t-1}\right) \implies \mathbb{E}(X_t \mid X_1^{t-1}) = \underbrace{\frac{N\mu - \sum_{i=1}^{t-1} X_i}{N - t + 1}}_{=: \, \mu_t^{\mathrm{WoR}}}$$

Consider a “candidate mean” $m \in [0,1]$

Consider a "candidate mean" $m \in [0,1]$

$K_0 \leftarrow \$1$

Consider a "candidate mean" $m \in [0,1]$

$K_0 \leftarrow \$1$

For $t = 1,2,3,\ldots$:

Consider a "candidate mean" $m \in [0,1]$

$K_0 \leftarrow \$1$

For $t = 1,2,3,\ldots$:

Gambler chooses bet $\lambda_t \in (-1/(1-m_t^{\text{WoR}}),\ 1/m_t^{\text{WoR}})$

$$m_t^{\text{WoR}} = \frac{Nm - \sum_{i=1}^{t-1} X_i}{N - t + 1}$$

Consider a "candidate mean" $m \in [0,1]$

$K_0 \leftarrow \$1$

For $t = 1,2,3,\ldots$:

Gambler chooses bet $\lambda_t \in (-1/(1-m_t^{\text{WoR}}),\ 1/m_t^{\text{WoR}})$

Observe X_t

$$m_t^{\text{WoR}} = \frac{Nm - \sum_{i=1}^{t-1} X_i}{N - t + 1}$$

Consider a "candidate mean" $m \in [0,1]$

$K_0 \leftarrow \$1$

For $t = 1,2,3,\ldots$:

Gambler chooses bet $\lambda_t \in (-1/(1-m_t^{\mathrm{WoR}}),\ 1/m_t^{\mathrm{WoR}})$

Observe X_t

$K_t \leftarrow K_{t-1} \ + \ K_{t-1} \cdot \lambda_t \cdot (X_t - m_t^{\mathrm{WoR}})$

EndFor

$$m_t^{\mathrm{WoR}} = \frac{Nm - \sum_{i=1}^{t-1} X_i}{N - t + 1}$$

Consider a "candidate mean" $m \in [0,1]$

$K_0 \leftarrow \$1$

For $t = 1,2,3,\ldots$:

Gambler chooses bet $\lambda_t \in (-1/(1-m),\ 1/m)$

Observe X_t

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

EndFor

Consider a “candidate mean” $m \in [0,1]$

$K_0 \leftarrow \$1$

For $t = 1,2,3,\ldots$:

Gambler chooses bet $\lambda_t \in (-1/(1-m_t^{\text{WoR}}),\ 1/m_t^{\text{WoR}})$

Observe X_t

$K_t \leftarrow K_{t-1} \ + \ K_{t-1} \cdot \lambda_t \cdot (X_t - m_t^{\text{WoR}})$

EndFor

$$m_t^{\text{WoR}} = \frac{Nm - \sum_{i=1}^{t-1} X_i}{N - t + 1}$$

Then,

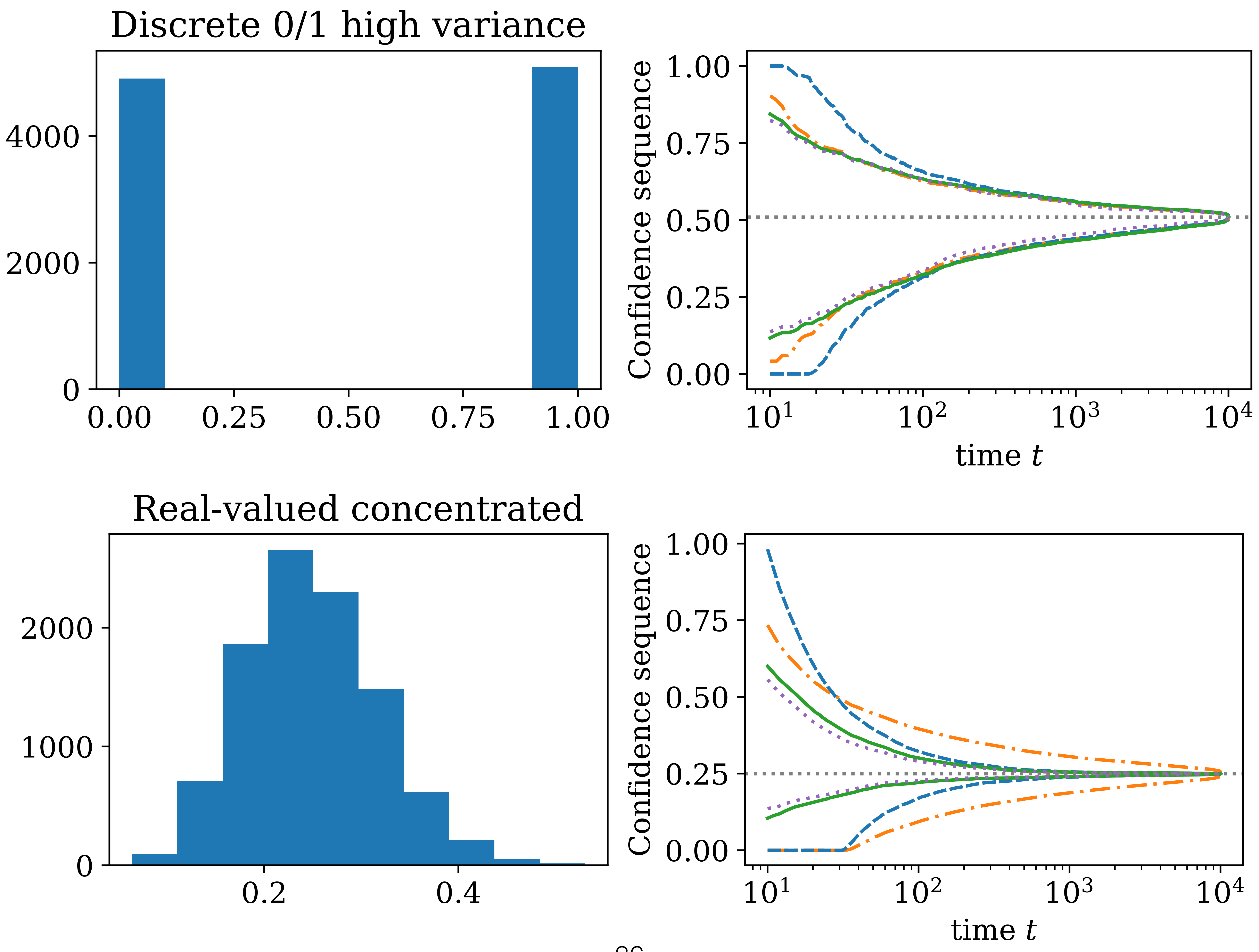
$$K_t^{\mathrm{WoR}}(\mu) := \prod_{i=1}^{t} (1 + \lambda_i \cdot (X_i - \mu_t^{\mathrm{WoR}}))$$

forms a nonnegative martingale, and

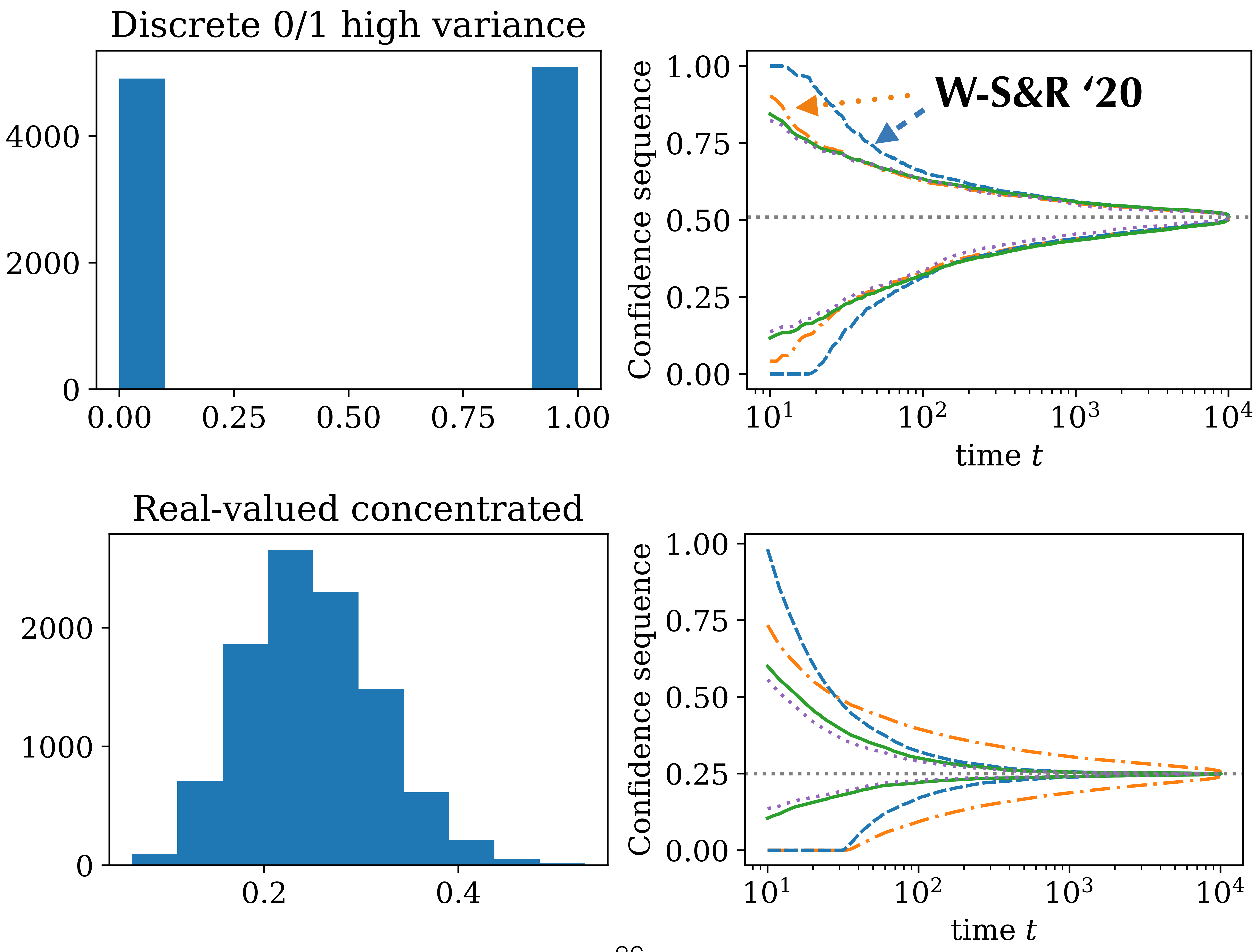
$$C_t^{\mathrm{WoR}} := \left\{ m \in [0,1] : K_t^{\mathrm{WoR}}(m) < \frac{1}{\alpha} \right\}$$

forms a $(1 - \alpha)$-confidence sequence.

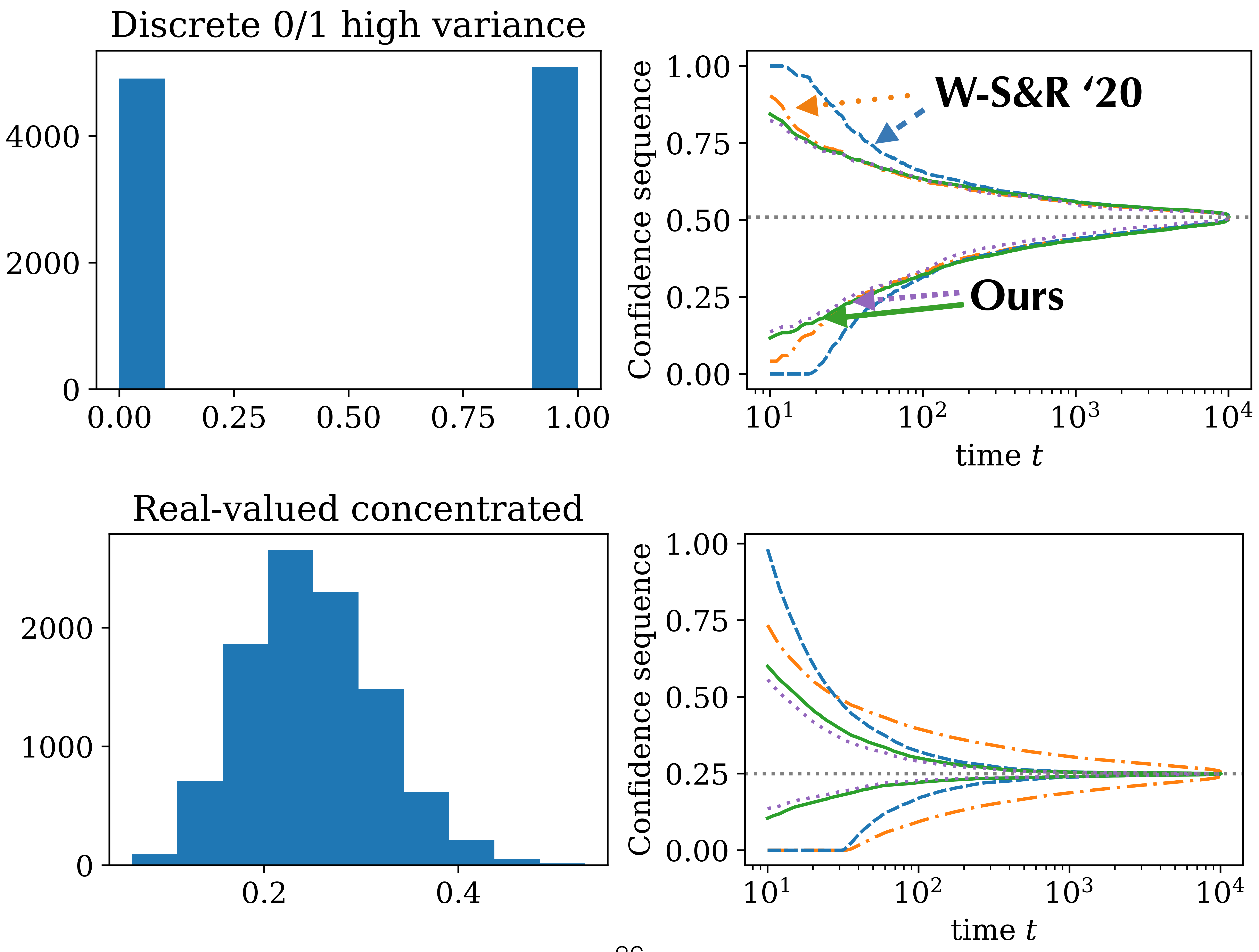
Confidence sequences for sampling WoR



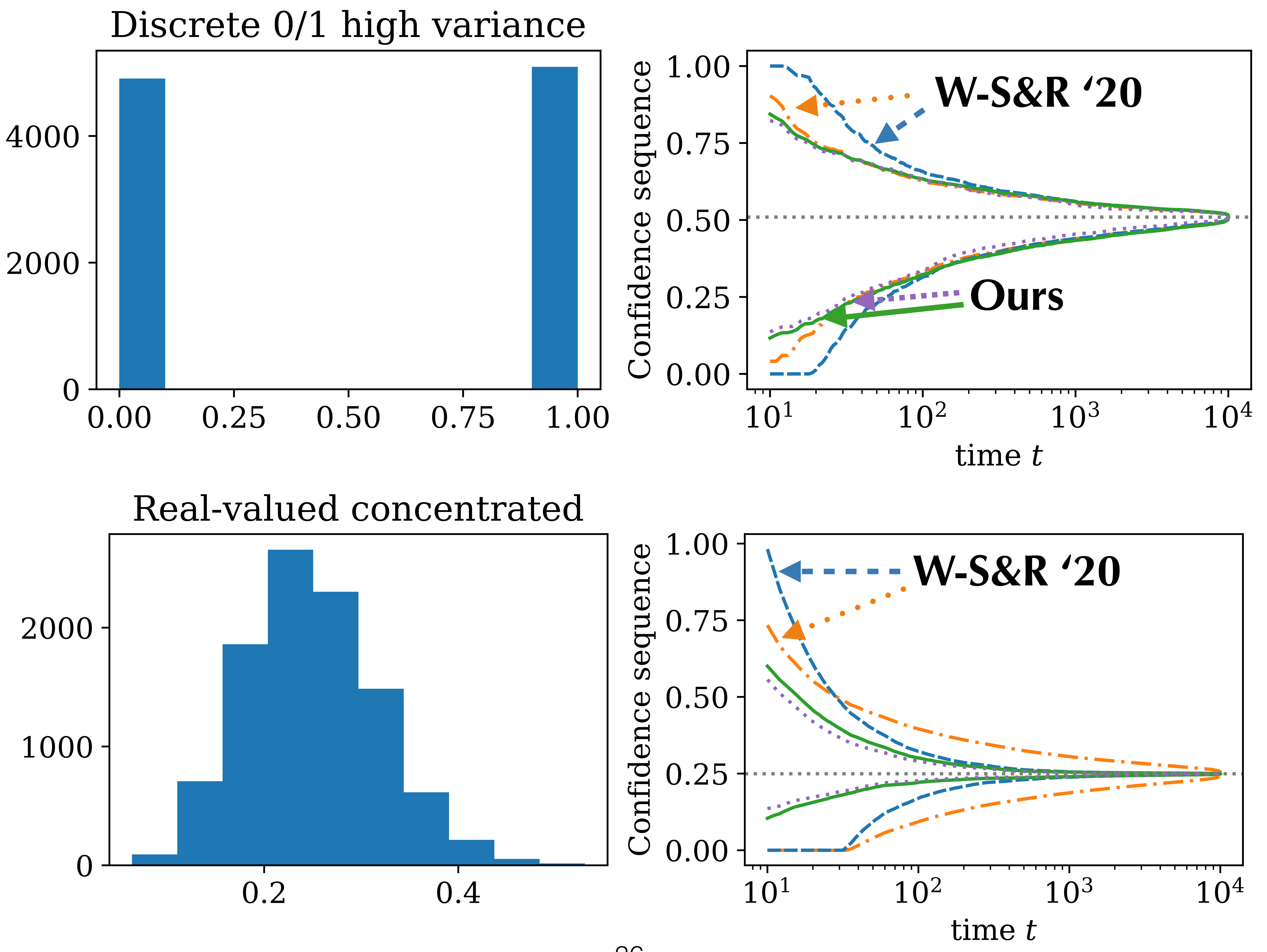
Confidence sequences for sampling WoR



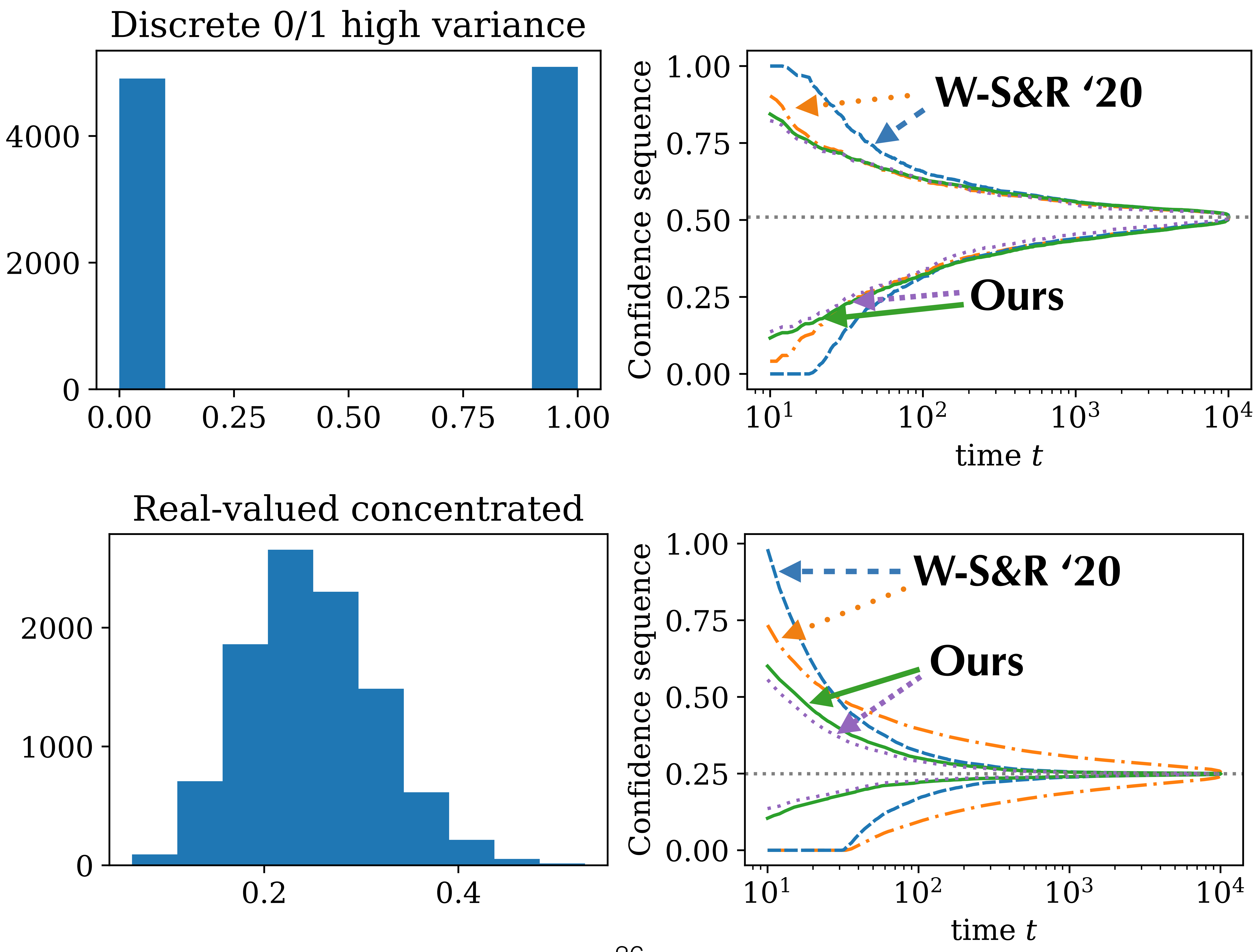
Confidence sequences for sampling WoR



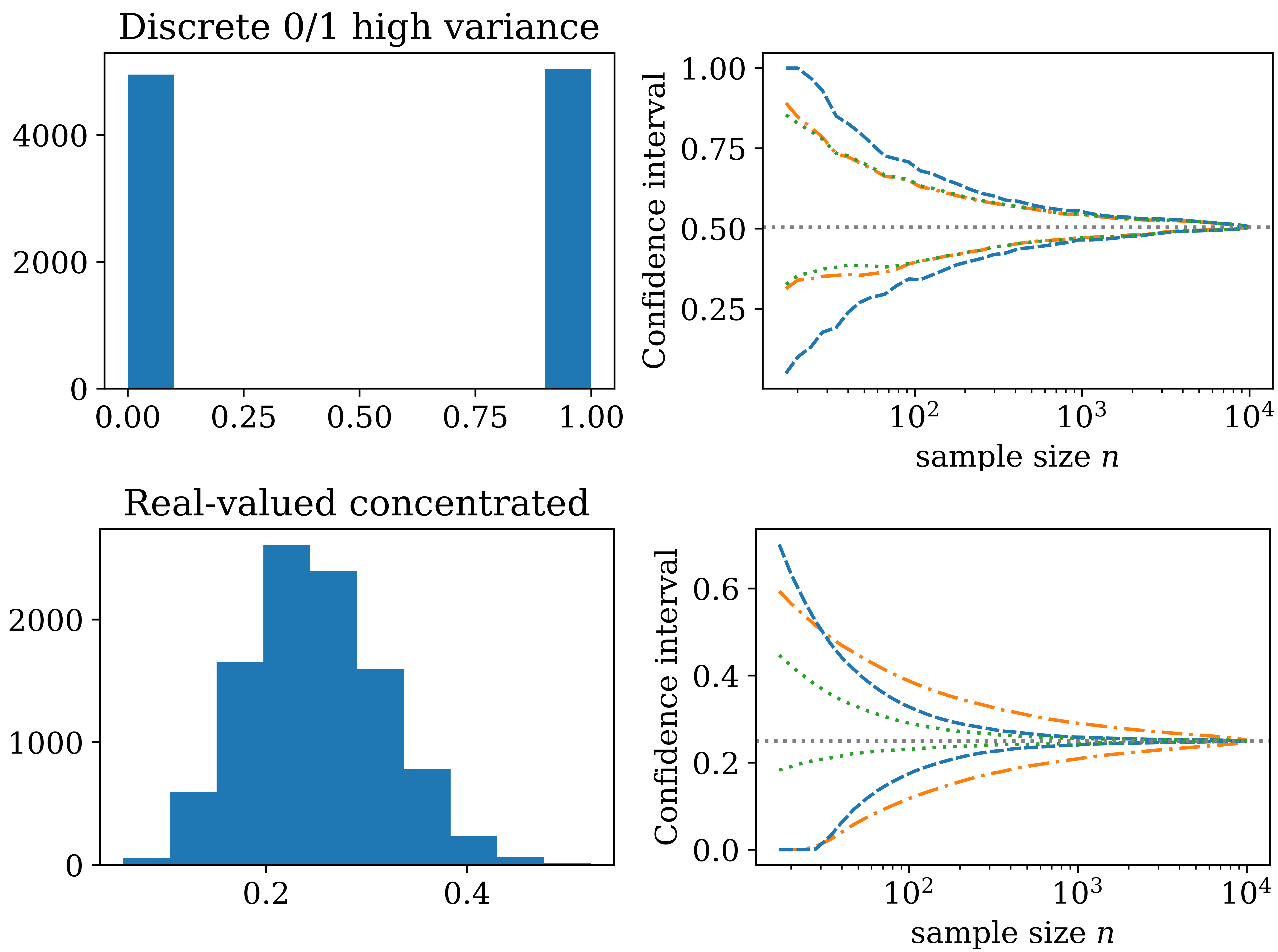
Confidence sequences for sampling WoR



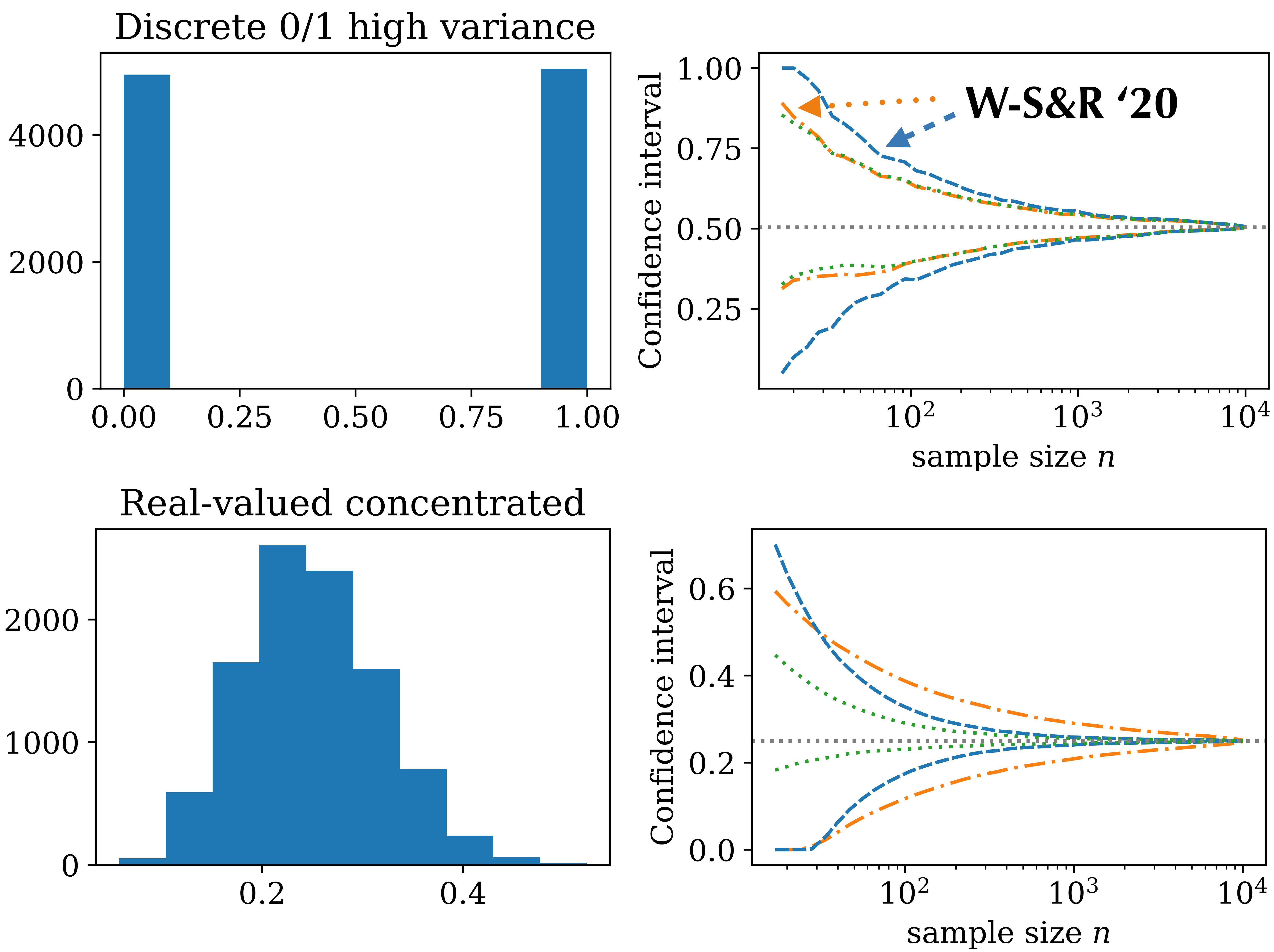
Confidence sequences for sampling WoR



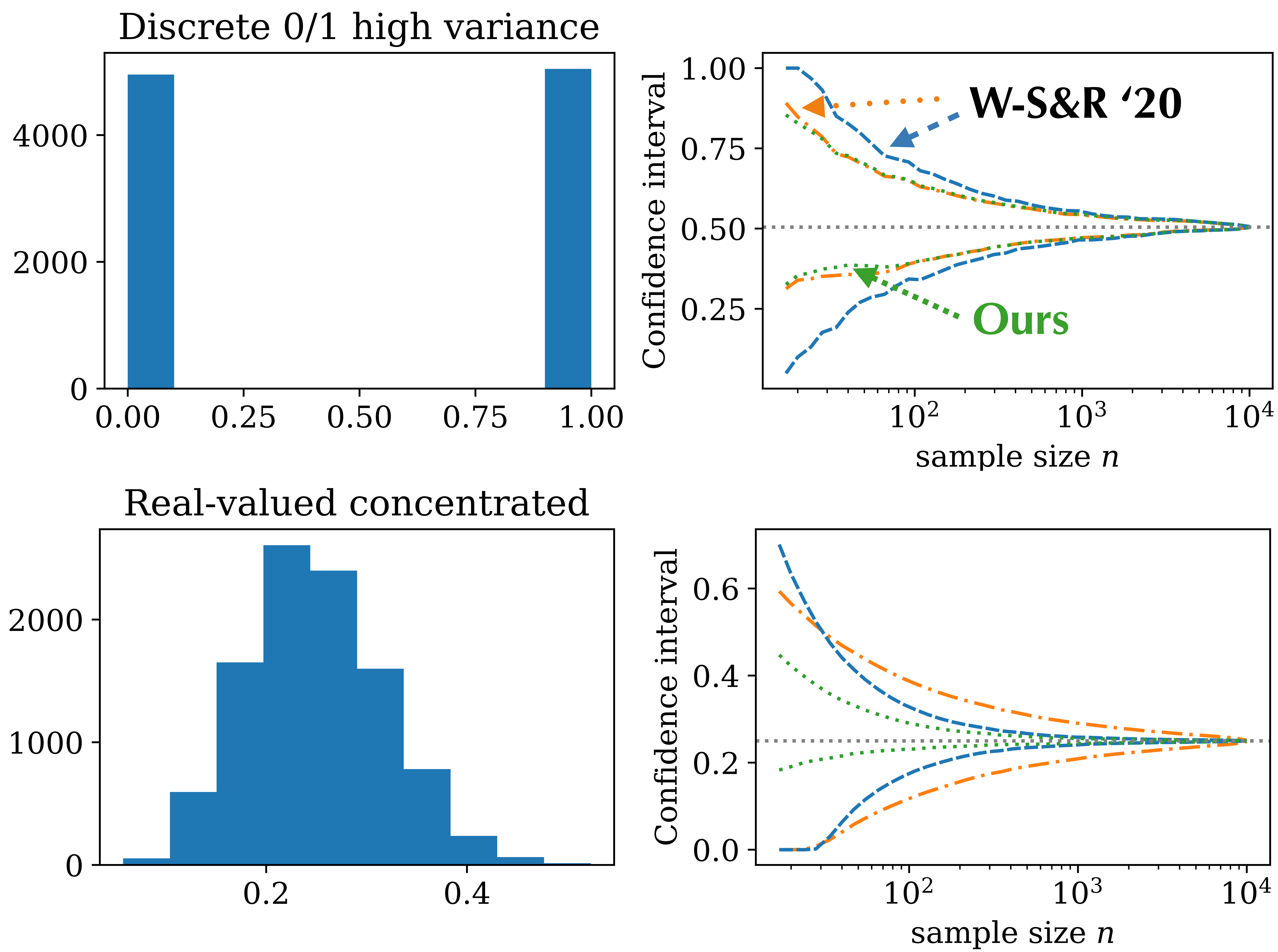
Confidence intervals for sampling WoR



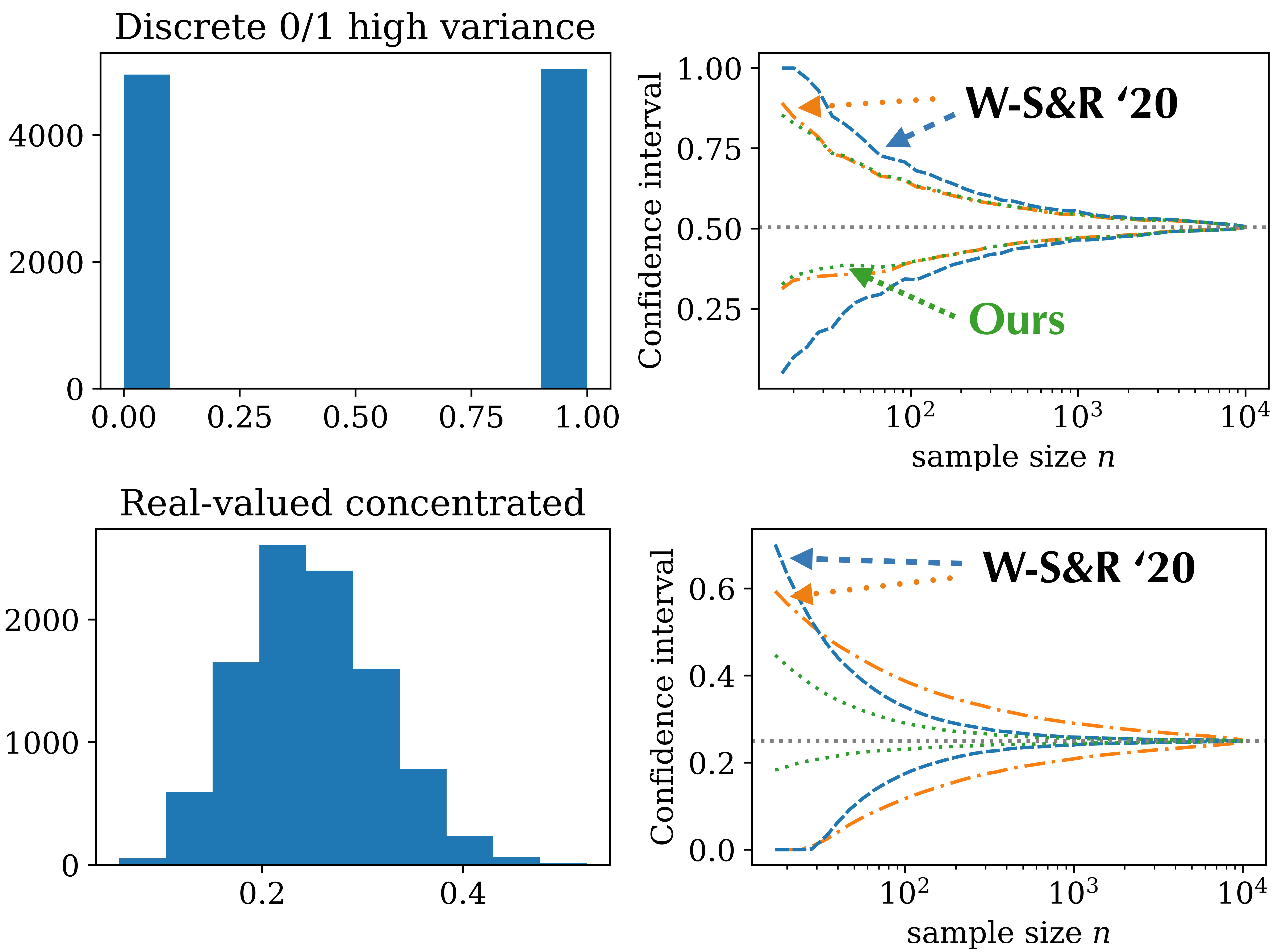
Confidence intervals for sampling WoR



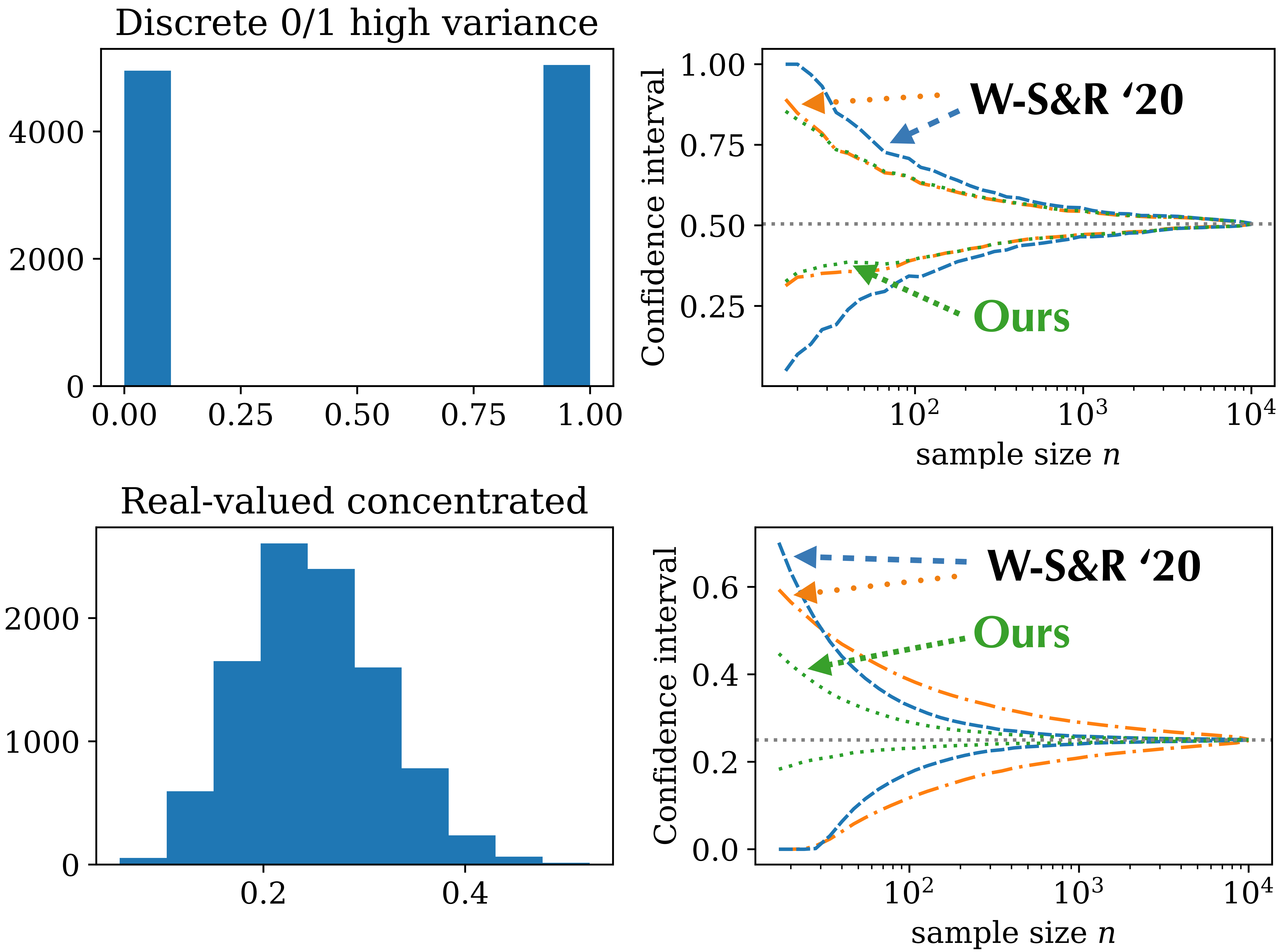
Confidence intervals for sampling WoR



Confidence intervals for sampling WoR



Confidence intervals for sampling WoR



Closed-form empirical Bernstein confidence sequences & confidence intervals

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

While C_t is easy to compute, it is not closed-form.

However,

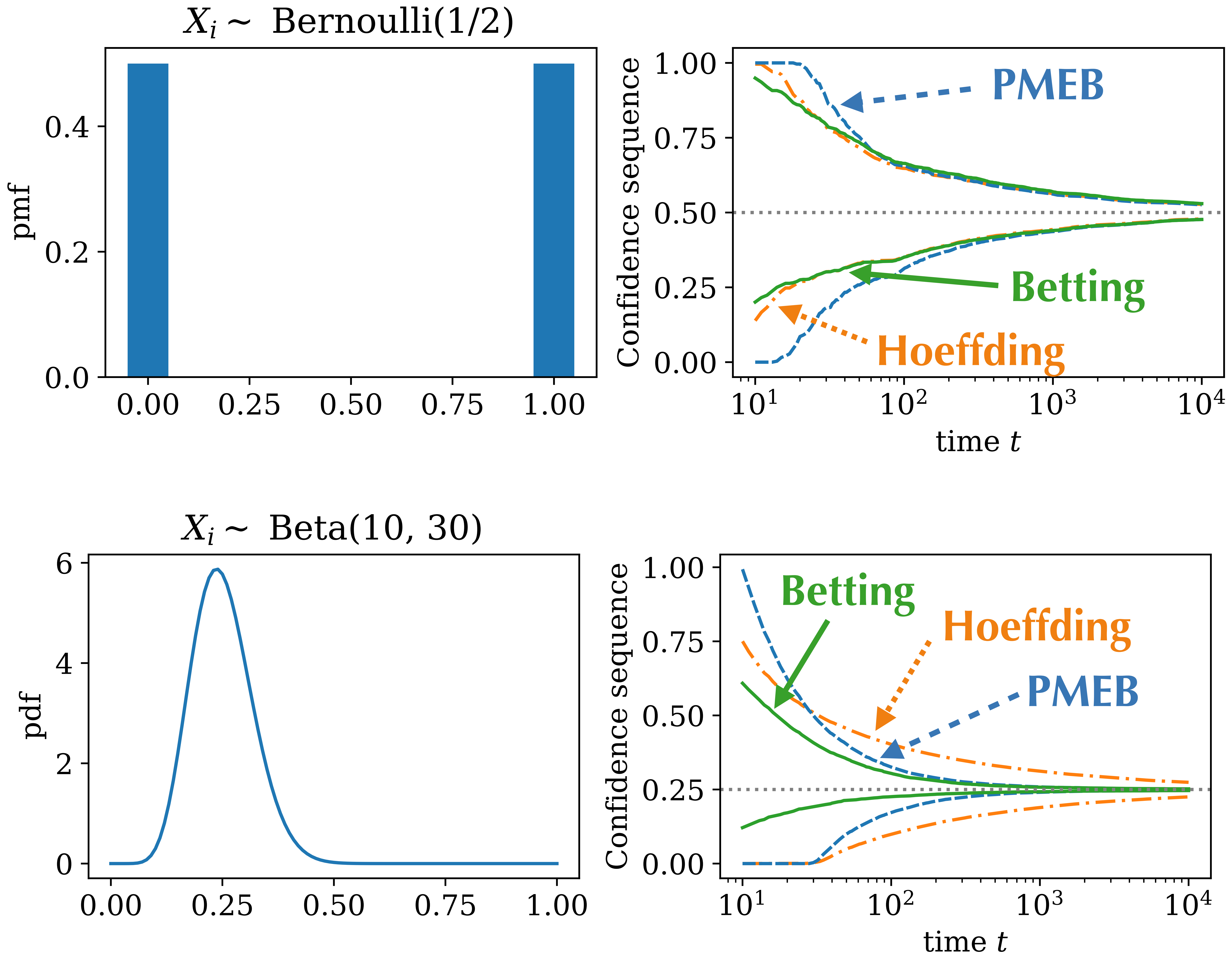
$$C_t^{\text{PMEB}} := \left\{ m \in [0,1] : \prod_{i=1}^{t} \exp\left\{ \lambda_i(X_i - m) - 4(X_i - \widehat{\mu}_{i-1})^2 \psi_E(\lambda_i) \right\} \right\}$$ **is!**

$$C_t^{\mathrm{PMEB}} := \left(\frac{\sum_{i=1}^t \lambda_i X_i}{\sum_{i=1}^t \lambda_i} \pm \frac{\log(2/\alpha) + 4\sum_{i=1}^t (X_i - \widehat{\mu}_{i-1})^2 \psi_E(\lambda_i)}{\sum_{i=1}^t \lambda_i} \right)$$

where $\quad \psi_E(\lambda) := -(\log(1-\lambda) - \lambda)/4,$

$$\lambda_t := \sqrt{\frac{2\log(2/\alpha)}{\widehat{\sigma}_{t-1} t \log(1+t)}} \wedge \frac{1}{2},$$

$$\widehat{\sigma}_t^2 := \frac{1/4 + \sum_{i=1}^t (X_i - \widehat{\mu}_i^2)^2}{t+1}, \quad \text{and} \quad \widehat{\mu}_t := \frac{1/2 + \sum_{i=1}^t X_i}{t+1}.$$

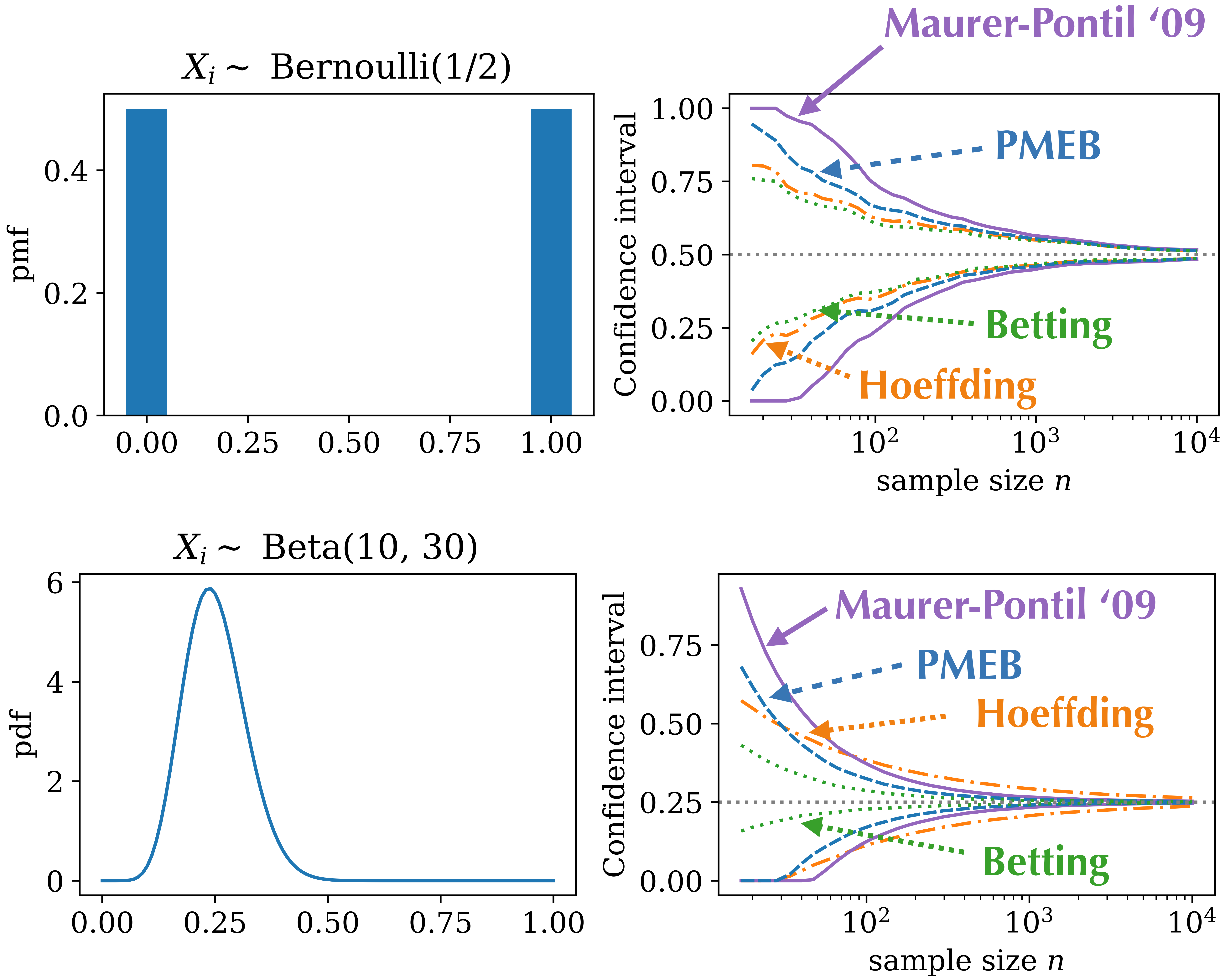

$X_i \sim$ Bernoulli(1/2)
pmf
0.0
0.2
0.4
0.00
0.25
0.50
0.75
1.00
Confidence sequence
1.00
0.75
0.50
0.25
0.00
PMEB
Betting
Hoeffding
10^1
10^2
10^3
10^4
time t
$X_i \sim$ Beta(10, 30)
pdf
6
4
2
0
0.00
0.25
0.50
0.75
1.00
Confidence sequence
1.00
0.75
0.50
0.25
0.00
Betting
Hoeffding
PMEB
10^1
10^2
10^3
10^4
time t

Similarly for fixed-time confidence intervals:

$$C_n^{\mathrm{PMEB}} := \left(\frac{\sum_{i=1}^{n} \lambda_i X_i}{\sum_{i=1}^{n} \lambda_i} \pm \frac{\log(2/\alpha) + 4 \sum_{i=1}^{n} (X_i - \widehat{\mu}_{i-1})^2 \psi_E(\lambda_i)}{\sum_{i=1}^{n} \lambda_i} \right)$$

but here, $\lambda_i := \sqrt{\dfrac{2 \log(2/\alpha)}{n \widehat{\sigma}_{i-1}}} \wedge \dfrac{1}{2},$

Final bound: $\bigcap_{i \leq n} C_i^{\mathrm{PMEB}}$


$X_i \sim$ Bernoulli(1/2)
pmf
Maurer-Pontil '09
PMEB
Betting
Hoeffding
Confidence interval
sample size n
$X_i \sim$ Beta(10, 30)
pdf
Maurer-Pontil '09
PMEB
Hoeffding
Betting
Confidence interval
sample size n

Choice of $(\lambda_t^+)_{t=1}^n$ and $(\lambda_t^-)_{t=1}^n$ for fixed-time confidence intervals

Why does perform so well?

$$K_n^+(\mu) := \prod_{i=1}^{n} \left(1 + \lambda \cdot (X_i - \mu)\right)$$

$$\gtrsim \prod_{i=1}^{n} \exp\left\{\lambda \cdot (X_i - \mu) - (X_i - \widehat{\mu}_{i-1})^2 \lambda^2 / 2\right\}$$

$$\implies \text{Width}_n := \frac{\log(2/\alpha) + \sum_{i=1}^{n} (X_i - \widehat{\mu}_{i-1})^2/2}{t\lambda}$$

$$\approx \frac{\log(2/\alpha) + n\sigma^2/2}{t\lambda}$$

Why does $\lambda_i^+(m) := \sqrt{\frac{2\log(2/\alpha)}{n\widehat{\sigma}_{i-1}^2}} \wedge \frac{1}{m}$ perform so well?

$$
\begin{aligned}
K_n^+(\mu) &:= \prod_{i=1}^{n} \left(1 + \lambda \cdot (X_i - \mu)\right) \\
&\gtrsim \underbrace{\prod_{i=1}^{n} \exp\left\{\lambda \cdot (X_i - \mu) - (X_i - \widehat{\mu}_{i-1})^2 \lambda^2/2\right\}}
\end{aligned}
$$

$$
\begin{aligned}
\implies \text{Width}_n &:= \frac{\log(2/\alpha) + \sum_{i=1}^n (X_i - \widehat{\mu}_{i-1})^2/2}{t\lambda} \\
&\approx \frac{\log(2/\alpha) + n\sigma^2/2}{t\lambda}
\end{aligned}
$$

$$\underset{\lambda}{\operatorname{argmin}}\ \mathrm{Width}_n = \sqrt{\frac{2\log(2/\alpha)}{n\sigma^2}}$$

$$\underset{\lambda}{\operatorname{argmin}}\ \mathrm{Width}_n = \sqrt{\frac{2\log(2/\alpha)}{n\sigma^2}}$$

$$\lambda_t^+(m) := \sqrt{\frac{2\log(2/\alpha)}{n\,\widehat{\sigma}_{t-1}^2}} \wedge \frac{1}{m}$$

$$\underset{\lambda}{\operatorname{argmin}}\ \mathrm{Width}_n = \sqrt{\frac{2\log(2/\alpha)}{n\sigma^2}}$$

$$\lambda_t^+(m) := \sqrt{\frac{2\log(2/\alpha)}{n\,\widehat{\sigma}_{t-1}^2}} \wedge \frac{1}{m} \qquad \lambda_t^-(m) := \sqrt{\frac{2\log(2/\alpha)}{n\,\widehat{\sigma}_{t-1}^2}} \wedge \frac{1}{1-m}$$

Brief selective history of betting ideas