Uniform, nonasymptotic confidence sequences for sequential treatment effect estimation

Steve Howard
Joint work with Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon

August 10, 2018

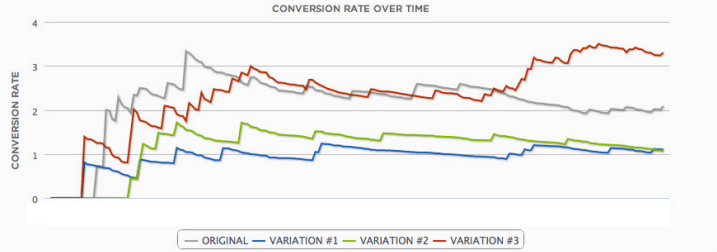
Sequential monitoring of experiment results is problematic.

Click on the button

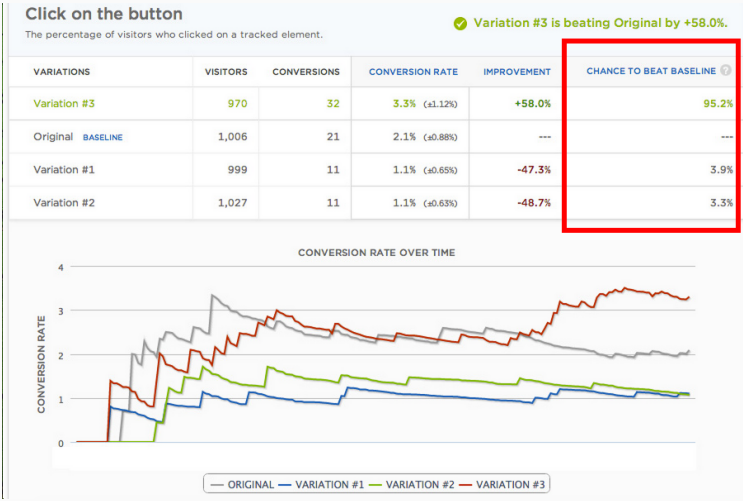
The percentage of visitors who clicked on a tracked element.

✓ Variation #3 is beating Original by +58.0%.

VARIATIONS	VISITORS	CONVERSIONS	CONVERSION RATE	IMPROVEMENT	CHANCE TO BEAT BASELINE
Variation #3	970	32	3.3% (±1.12%)	+58.0%	95.2%
Original BASELINE	1,006	21	2.1% (±0.88%)	---	---
Variation #1	999	11	1.1% (±0.65%)	-47.3%	3.9%
Variation #2	1,027	11	1.1% (±0.63%)	-48.7%	3.3%



Sequential monitoring of experiment results is problematic.



Click on the button

The percentage of visitors who clicked on a tracked element.

Variation #3 is beating Original by +58.0%.

VARIATIONS	VISITORS	CONVERSIONS	CONVERSION RATE	IMPROVEMENT	CHANCE TO BEAT BASELINE
Variation #3	970	32	3.3% (±1.12%)	+58.0%	95.2%
Original BASELINE	1,006	21	2.1% (±0.88%)	---	---
Variation #1	999	11	1.1% (±0.65%)	-47.3%	3.9%
Variation #2	1,027	11	1.1% (±0.63%)	-48.7%	3.3%

(Optimizely actually does the right thing now. [Johari et al. 2015])

We give a method to

- sequentially gather data from a randomized experiment,

We give a method to

- sequentially gather data from a randomized experiment,
- repeatedly compute a confidence interval for average treatment effect,

We give a method to

- sequentially gather data from a randomized experiment,
- repeatedly compute a confidence interval for average treatment effect,
- and stop the experiment at any time,

We give a method to

- sequentially gather data from a randomized experiment,
- repeatedly compute a confidence interval for average treatment effect,
- and stop the experiment at any time,

with the following guarantee:

No matter how you adaptively stop, you have a valid confidence interval for the average treatment effect.

We give a method to

- sequentially gather data from a randomized experiment,
- repeatedly compute a confidence interval for average treatment effect,
- and stop the experiment at any time,

with the following guarantee:

No matter how you adaptively stop, you have a valid confidence interval for the average treatment effect.

- No superpopulation or stationarity assumptions.
- No knowledge of the stopping rule.
- No bound on sample size.
- No asymptotic approximations or sharp null hypothesis.

We assume fixed potential outcomes and bounded individual treatment effects.

Unit i has fixed potential outcomes $Y_i(0), Y_i(1)$, for $i = 1, 2, \ldots$

We assume fixed potential outcomes and bounded individual treatment effects.

Unit i has fixed potential outcomes $Y_i(0), Y_i(1)$, for $i = 1, 2, \ldots$

Assign random treatment $T_i \in \{0, 1\}$ to individual i.
Observe $Y_i(T_i)$. [Neyman 1923, Rubin 1974]

We assume fixed potential outcomes and bounded individual treatment effects.

Unit i has fixed potential outcomes $Y_i(0), Y_i(1)$, for $i = 1, 2, \ldots$

Assign random treatment $T_i \in \{0, 1\}$ to individual i.
Observe $Y_i(T_i)$. [Neyman 1923, Rubin 1974]

Assumption: $|Y_i(1) - Y_i(0)| \leq \tau^{\max}$ for all i.

We assume fixed potential outcomes and bounded individual treatment effects.

Unit i has fixed potential outcomes $Y_i(0), Y_i(1)$, for $i = 1, 2, \ldots$

Assign random treatment $T_i \in \{0, 1\}$ to individual i.
Observe $Y_i(T_i)$. [Neyman 1923, Rubin 1974]

Assumption: $|Y_i(1) - Y_i(0)| \leq \tau^{\max}$ for all i.

- Without some such assumption, impossible to construct a confidence interval
- One unit can have arbitrarily large influence

We assume fixed potential outcomes and bounded individual treatment effects.

Unit i has fixed potential outcomes $Y_i(0), Y_i(1)$, for $i = 1, 2, \ldots$

Assign random treatment $T_i \in \{0, 1\}$ to individual i.
Observe $Y_i(T_i)$. [Neyman 1923, Rubin 1974]

Assumption: $|Y_i(1) - Y_i(0)| \leq \tau^{\max}$ for all i.

- Without some such assumption, impossible to construct a confidence interval
- One unit can have arbitrarily large influence

For simplicity, in this talk, we'll assume $\mathbb{P}(T_i) = 1/2$ for all i.

We define a sequence of average treatment effect estimands.

Our goal: after observing units $1, \ldots, t$, we'd like to estimate

$$\text{ATE}_t := \frac{1}{t} \sum_{i=1}^{t} [Y_i(1) - Y_i(0)].$$

We define a sequence of average treatment effect estimands.

Our goal: after observing units $1, \ldots, t$, we'd like to estimate

$$\mathrm{ATE}_t := \frac{1}{t} \sum_{i=1}^{t} [Y_i(1) - Y_i(0)].$$

A **confidence sequence** for $(\mathrm{ATE}_t)_{t=1}^{\infty}$ is a sequence of intervals $(\mathrm{CI}_t)_{t=1}^{\infty}$ satisfying

$$\mathbb{P}(\mathrm{ATE}_t \in \mathrm{CI}_t \text{ for all } t \in \mathbb{N}) \geq 1 - \alpha.$$

[Darling and Robbins 1967, Lai 1984, Jennison and Turnbull 1989, Johari et al. 2015]

We define a sequence of average treatment effect estimands.

Our goal: after observing units $1, \ldots, t$, we'd like to estimate

$$\mathrm{ATE}_t := \frac{1}{t} \sum_{i=1}^{t} [Y_i(1) - Y_i(0)].$$

A **confidence sequence** for $(\mathrm{ATE}_t)_{t=1}^{\infty}$ is a sequence of intervals $(\mathrm{CI}_t)_{t=1}^{\infty}$ satisfying

$$\mathbb{P}(\mathrm{ATE}_t \in \mathrm{CI}_t \text{ for all } t \in \mathbb{N}) \geq 1 - \alpha.$$

[Darling and Robbins 1967, Lai 1984, Jennison and Turnbull 1989, Johari et al. 2015]

- Dual to sequential hypothesis tests, always-valid p-values

We rely on a Horvitz-Thompson estimator

We require predictions $\hat{Y}_i(1)$ and $\hat{Y}_i(0)$ for $Y_i(1)$ and $Y_i(0)$.

We rely on a Horvitz-Thompson estimator

We require predictions $\hat{Y}_i(1)$ and $\hat{Y}_i(0)$ for $Y_i(1)$ and $Y_i(0)$.

We construct an unbiased estimator of $Y_i(1) - Y_i(0)$:

$$X_i := 2\overbrace{(2T_i - 1)}^{\pm 1}\left(\underbrace{Y_i(T_i)}_{\text{observed outcome}} - \frac{\overbrace{\hat{Y}_i(1) + \hat{Y}_i(0)}^{\text{average prediction}}}{2}\right)$$

[Horvitz and Thompson 1952, Aronow and Middleton 2013]

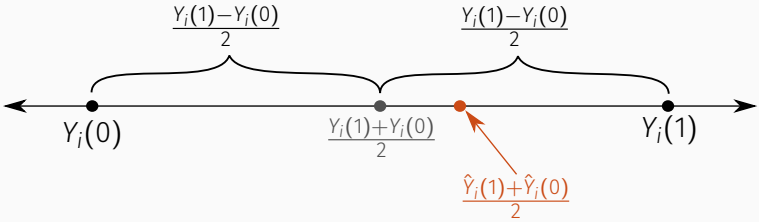
We rely on a Horvitz-Thompson estimator

We require predictions $\hat{Y}_i(1)$ and $\hat{Y}_i(0)$ for $Y_i(1)$ and $Y_i(0)$.

We construct an unbiased estimator of $Y_i(1) - Y_i(0)$:

$$X_i := \overbrace{2(2T_i - 1)}^{\pm 1} \left(\underbrace{Y_i(T_i)}_{\text{observed outcome}} - \overbrace{\frac{\hat{Y}_i(1) + \hat{Y}_i(0)}{2}}^{\text{average prediction}} \right)$$

[Horvitz and Thompson 1952, Aronow and Middleton 2013]



Estimation accuracy depends on prediction accuracy

$\sum_{i=1}^{t} X_i$ is sub-Gaussian with variance parameter $\sum_{i=1}^{t} \operatorname{Var} X_i$. We can apply a uniform Hoeffding bound. [Darling and Robbins 1967, 1968; Balsubramani 2014; Jamieson et al. 2014; Kaufmann et al. 2014; Balsubramani and Ramdas 2016; Zhao et al. 2016; Tchetgen Tchetgen and VanderWeele 2012]

Estimation accuracy depends on prediction accuracy

$\sum_{i=1}^{t} X_i$ is sub-Gaussian with variance parameter $\sum_{i=1}^{t} \operatorname{Var} X_i$. We can apply a uniform Hoeffding bound. [Darling and Robbins 1967, 1968; Balsubramani 2014; Jamieson et al. 2014; Kaufmann et al. 2014; Balsubramani and Ramdas 2016; Zhao et al. 2016; Tchetgen Tchetgen and VanderWeele 2012]

The variance of our estimate depends on unobserved prediction error:

$$\operatorname{Var} X_i = 4 \left(\frac{Y_i(1) + Y_i(0)}{2} - \frac{\hat{Y}_i(1) + \hat{Y}_i(0)}{2} \right)^2 .$$

Estimation accuracy depends on prediction accuracy

$\sum_{i=1}^{t} X_i$ is sub-Gaussian with variance parameter $\sum_{i=1}^{t} \operatorname{Var} X_i$. We can apply a uniform Hoeffding bound. [Darling and Robbins 1967, 1968; Balsubramani 2014; Jamieson et al. 2014; Kaufmann et al. 2014; Balsubramani and Ramdas 2016; Zhao et al. 2016; Tchetgen Tchetgen and VanderWeele 2012]

The variance of our estimate depends on unobserved prediction error:

$$\operatorname{Var} X_i = 4 \left(\frac{Y_i(1) + Y_i(0)}{2} - \frac{\hat{Y}_i(1) + \hat{Y}_i(0)}{2} \right)^2 .$$

We estimate $\operatorname{Var} X_i$ using observed prediction error

$$\widehat{\operatorname{Var}} X_i := 4[Y_i(T_i) - \hat{Y}_i(T_i)]^2.$$

Estimation accuracy depends on prediction accuracy

$\sum_{i=1}^{t} X_i$ is sub-Gaussian with variance parameter $\sum_{i=1}^{t} \operatorname{Var} X_i$. We can apply a uniform Hoeffding bound. [Darling and Robbins 1967, 1968; Balsubramani 2014; Jamieson et al. 2014; Kaufmann et al. 2014; Balsubramani and Ramdas 2016; Zhao et al. 2016; Tchetgen Tchetgen and VanderWeele 2012]

The variance of our estimate depends on unobserved prediction error:

$$\operatorname{Var} X_i = 4 \left(\frac{Y_i(1) + Y_i(0)}{2} - \frac{\hat{Y}_i(1) + \hat{Y}_i(0)}{2} \right)^2 .$$

We estimate $\operatorname{Var} X_i$ using observed prediction error

$$\widehat{\operatorname{Var}} X_i := 4[Y_i(T_i) - \hat{Y}_i(T_i)]^2.$$

This is conservatively biased.

We construct the confidence sequence using uniform sub-Gaussian concentration.

- $\bar{X}_t := t^{-1} \sum_{i=1}^{t} X_i$ is unbiased for ATE_t.

We construct the confidence sequence using uniform sub-Gaussian concentration.

- $\bar{X}_t \coloneqq t^{-1} \sum_{i=1}^{t} X_i$ is unbiased for ATE_t.
- $\hat{V}_t = 4t^{-2} \sum_{i=1}^{t} [Y_i(T_i) - \hat{Y}_i(T_i)]^2$ is conservatively biased for $\operatorname{Var} \bar{X}_t$.

We construct the confidence sequence using uniform sub-Gaussian concentration.

- $\bar{X}_t := t^{-1} \sum_{i=1}^{t} X_i$ is unbiased for ATE_t.
- $\hat{V}_t = 4t^{-2} \sum_{i=1}^{t} [Y_i(T_i) - \hat{Y}_i(T_i)]^2$ is conservatively biased for $\mathrm{Var}\, \bar{X}_t$.
- Use a uniform concentration bound to get an upper confidence sequence for $(\mathrm{Var}\, \bar{X}_t)_{t=1}^{\infty}$ in terms of $(\hat{V}_t)_{t=1}^{\infty}$.
  - This is where we use $\tau^{\max}$.

We construct the confidence sequence using uniform sub-Gaussian concentration.

- $\bar{X}_t := t^{-1} \sum_{i=1}^{t} X_i$ is unbiased for ATE_t.
- $\hat{V}_t = 4t^{-2} \sum_{i=1}^{t} [Y_i(T_i) - \hat{Y}_i(T_i)]^2$ is conservatively biased for $\mathrm{Var}\, \bar{X}_t$.
- Use a uniform concentration bound to get an upper confidence sequence for $(\mathrm{Var}\, \bar{X}_t)_{t=1}^{\infty}$ in terms of $(\hat{V}_t)_{t=1}^{\infty}$.
  - This is where we use $\tau^{\max}$.
- Use uniform concentration again to get a sequence of confidence radii $(R_t)_{t=1}^{\infty}$ satisfying

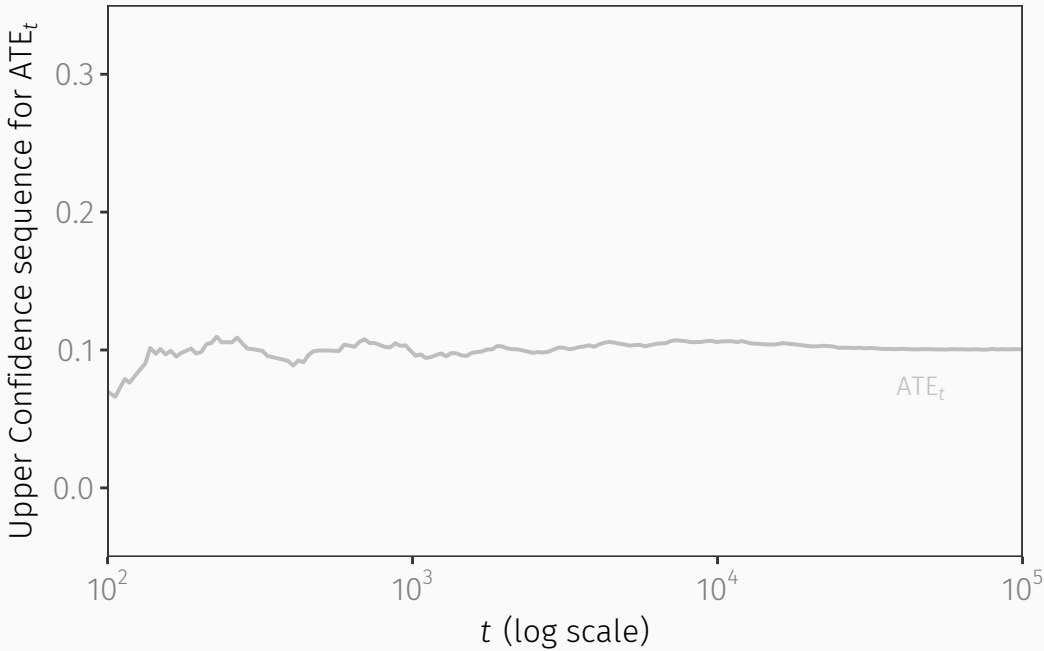
$$\mathbb{P}\left(\left|\bar{X}_t - \mathrm{ATE}_t\right| < R_t \text{ for all } t \in \mathbb{N}\right) \geq 1 - \alpha.$$

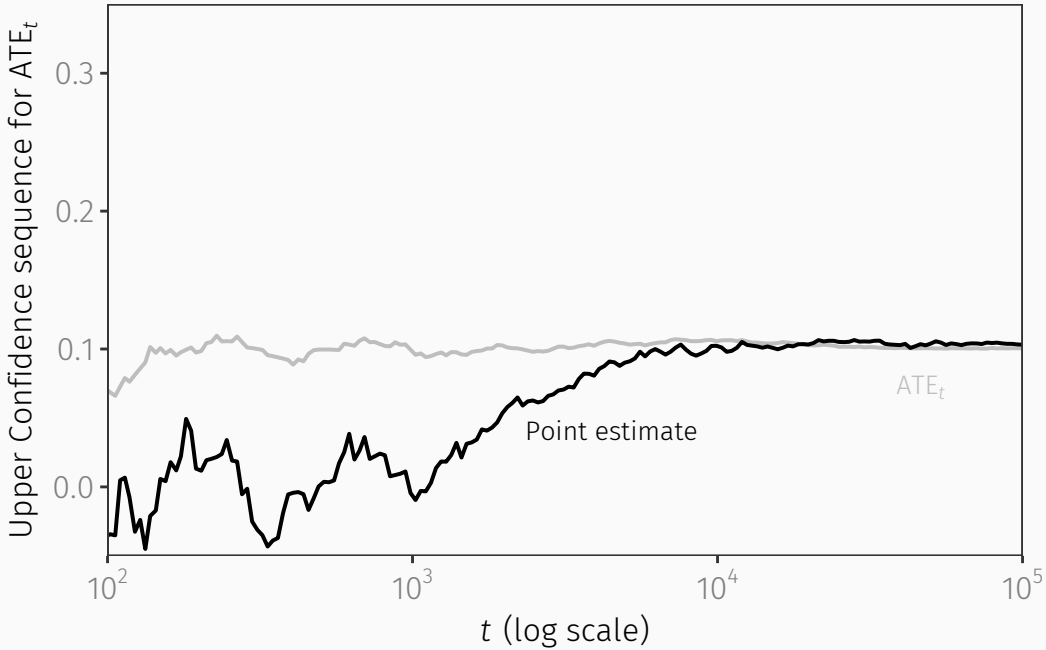
We construct the confidence sequence using uniform sub-Gaussian concentration.

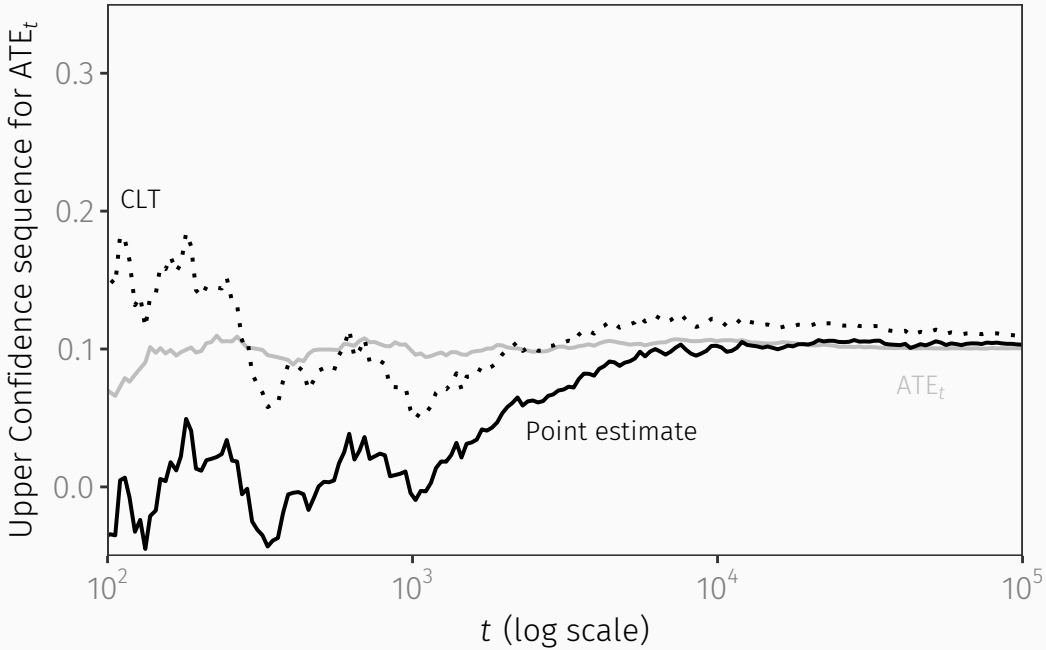
- $\bar{X}_t := t^{-1} \sum_{i=1}^{t} X_i$ is unbiased for ATE_t.
- $\hat{V}_t = 4t^{-2} \sum_{i=1}^{t} [Y_i(T_i) - \hat{Y}_i(T_i)]^2$ is conservatively biased for $\mathrm{Var}\, \bar{X}_t$.
- Use a uniform concentration bound to get an upper confidence sequence for $(\mathrm{Var}\, \bar{X}_t)_{t=1}^{\infty}$ in terms of $(\hat{V}_t)_{t=1}^{\infty}$.
  - This is where we use $\tau^{\max}$.
- Use uniform concentration again to get a sequence of confidence radii $(R_t)_{t=1}^{\infty}$ satisfying

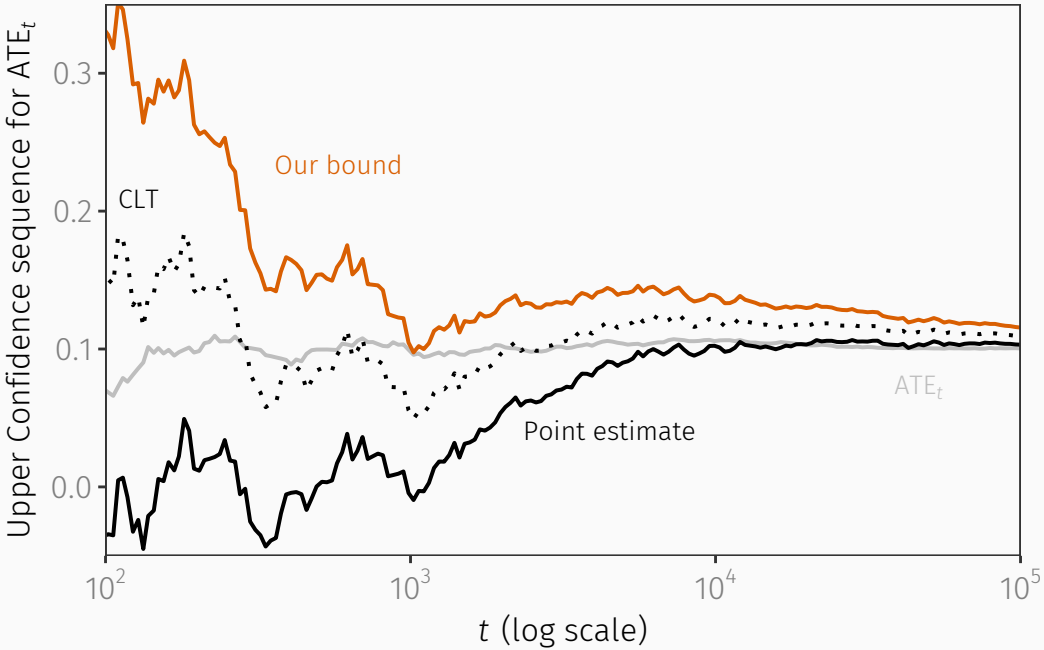
$$\mathbb{P}\left(\left|\bar{X}_t - \mathrm{ATE}_t\right| < R_t \text{ for all } t \in \mathbb{N}\right) \geq 1 - \alpha.$$

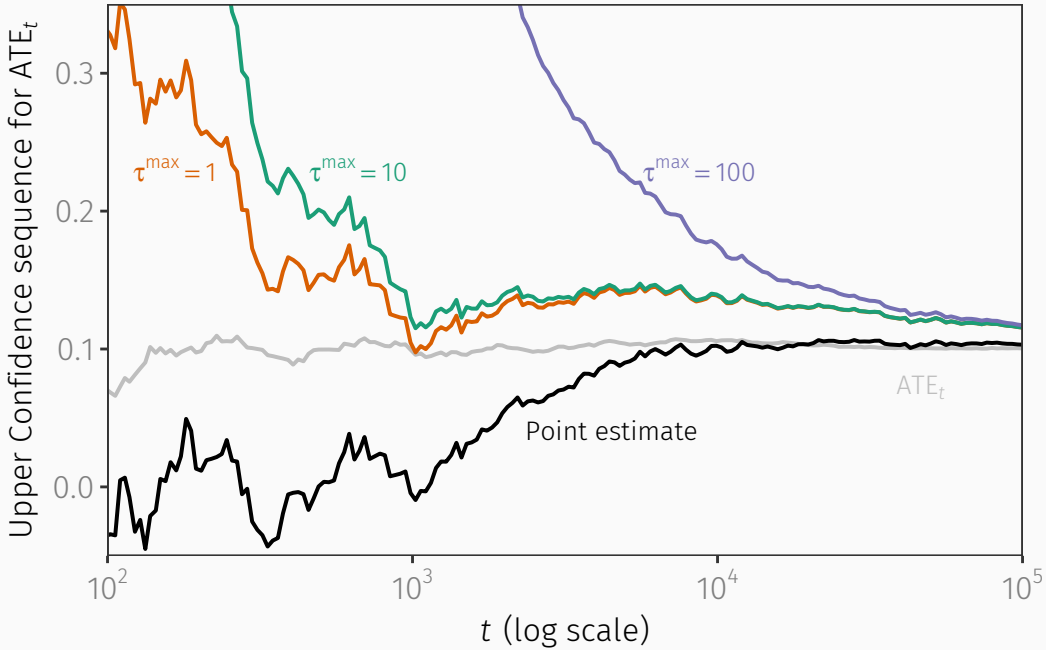
  - Typically, $R_t = \mathcal{O}_p\left(\sqrt{\frac{\log t}{t}}\right)$ or $\mathcal{O}_p\left(\sqrt{\frac{\log \log t}{t}}\right)$.
  - Large-sample decay rate of R_t does not depend on $\tau^{\max}$.


Upper Confidence sequence for ATE_t
0.3
0.2
0.1
0.0
ATE_t
10^2
10^3
10^4
10^5
t (log scale)


Upper Confidence sequence for ATE_t
0.3
0.2
0.1
0.0
ATE_t
Point estimate
10^2
10^3
10^4
10^5
t (log scale)


Upper Confidence sequence for ATE_t
0.3
0.2
0.1
0.0
CLT
ATE_t
Point estimate
10^2
10^3
10^4
10^5
t (log scale)


Upper Confidence sequence for ATE_t
0.3
0.2
0.1
0.0
Our bound
CLT
ATE_t
Point estimate
10^2
10^3
10^4
10^5
t (log scale)


Upper Confidence sequence for ATE_t
0.3
0.2
0.1
0.0
$\tau^{max} = 1$
$\tau^{max} = 10$
$\tau^{max} = 100$
ATE_t
Point estimate
10^2
10^3
10^4
10^5
t (log scale)

Recap

- Non-asymptotic confidence sequences for ATE_t
- Flexible inferential tool for sequential experiments
- Provable coverage under the assumption of bounded individual treatment effects
- Replace central limit theorem argument with uniform concentration bounds

Thank you!

stevehoward@berkeley.edu

Exponential line-crossing inequalities:
https://arxiv.org/abs/1808.03204