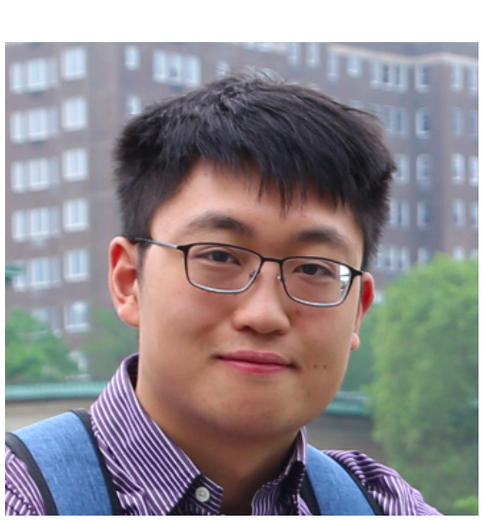
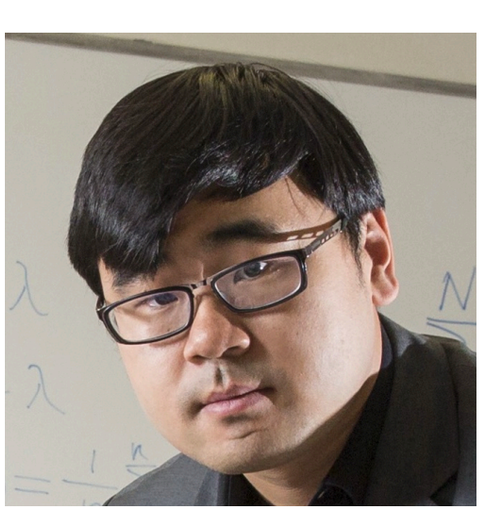
# Valid Inference under $S^3$ (selection, stopping, and [adaptive] sampling) bias for A/B testing

Twitter ML Modeling Seminar
July 13, 2022

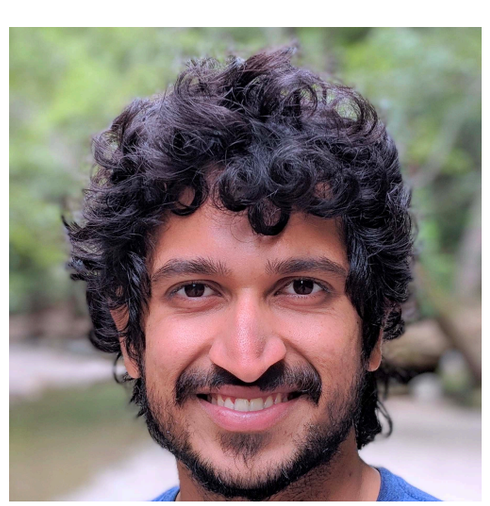
Joint work with:



Ziyu Xu (Neil)
(CMU)



Ruodu Wang
(Waterloo)



Aaditya Ramdas
(CMU)

Paper: Post-selection inference for e-value based confidence intervals [arXiv: 2203.12572]

UNIVERSITY OF WATERLOO



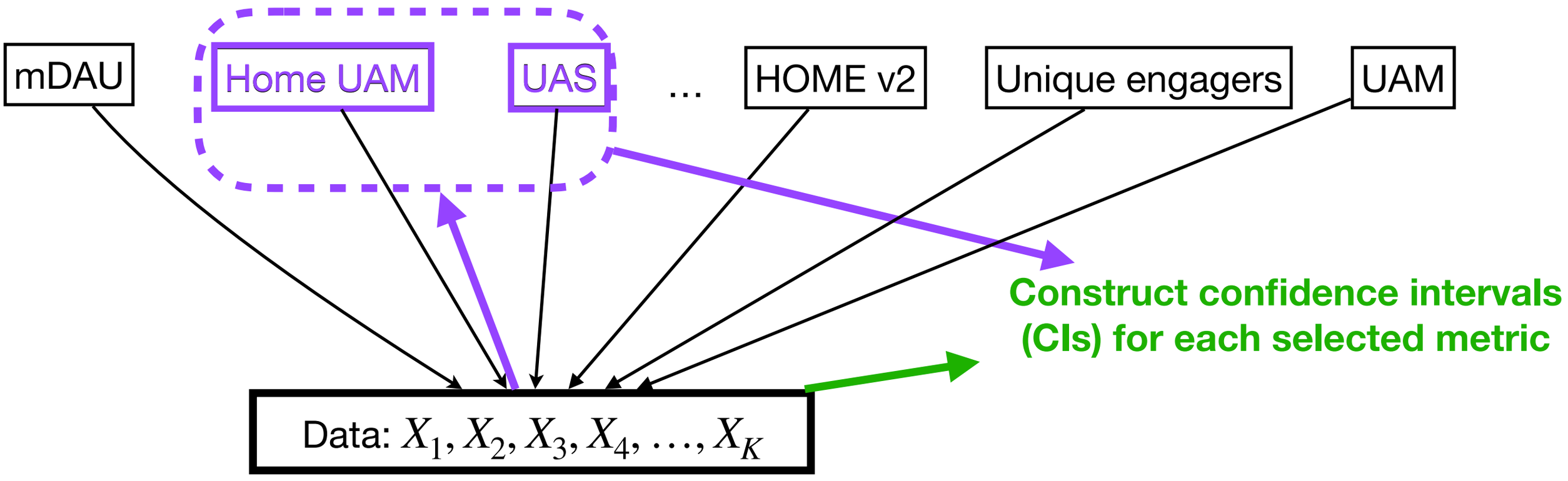
Carnegie Mellon University

# Motivating example: selecting metrics to justify shipping decisions

Initially, there may be $K$ metrics we wish to estimate.

For example, the following metrics from Home shipping criteria:

**Select ones with positive effect for estimation based on data**



# Outline

1. Introduce the post-selection inference problem.

2. Compare our method (**e-BY procedure**) to the current state-of-the-art (**BY procedure**).

3. Describe a novel category of confidence intervals: **e-CIs.**

4. Results of simulations in a nonparametric setting.

# Possible statistical guarantees under selection bias

$\theta_1^*, \ldots, \theta_K^*$ are the parameters we are initially interested in estimating.

**We have:** $C_i(\alpha)$ is the $(1-\alpha)$-CI we can construct for the $i$th parameter.

CI guarantee: $\mathbb{P}(\theta_i^* \in C_i(\alpha)) \geq 1 - \alpha$

Using some selection rule $\mathscr{R}$ to derive a selection set $\mathcal{S} = \mathscr{R}(\mathbf{X})$ of parameters to estimate induces a selection bias.

**We want:** Given any selected set $\mathscr{R}$, output CIs $C_1(\alpha_1), \ldots, C_K(\alpha_K)$ at corrected levels $\alpha_1, \ldots, \alpha_K$ such that we can maintain some form of statistical validity.

# False coverage rate (FCR): aggregate statistical validity

Define an empirical quantity: false coverage proportion (FCP).

$$\text{FCP} := \frac{\sum_{i \in \mathcal{S}} \mathbf{1}\{\theta_i^* \notin C_i(\alpha_i)\}}{|\mathcal{S}| \vee 1} \qquad \text{FCR} := \mathbb{E}[\text{FCP}]$$

FCR is an aggregate measure of false coverage
across the CIs of selected parameters.

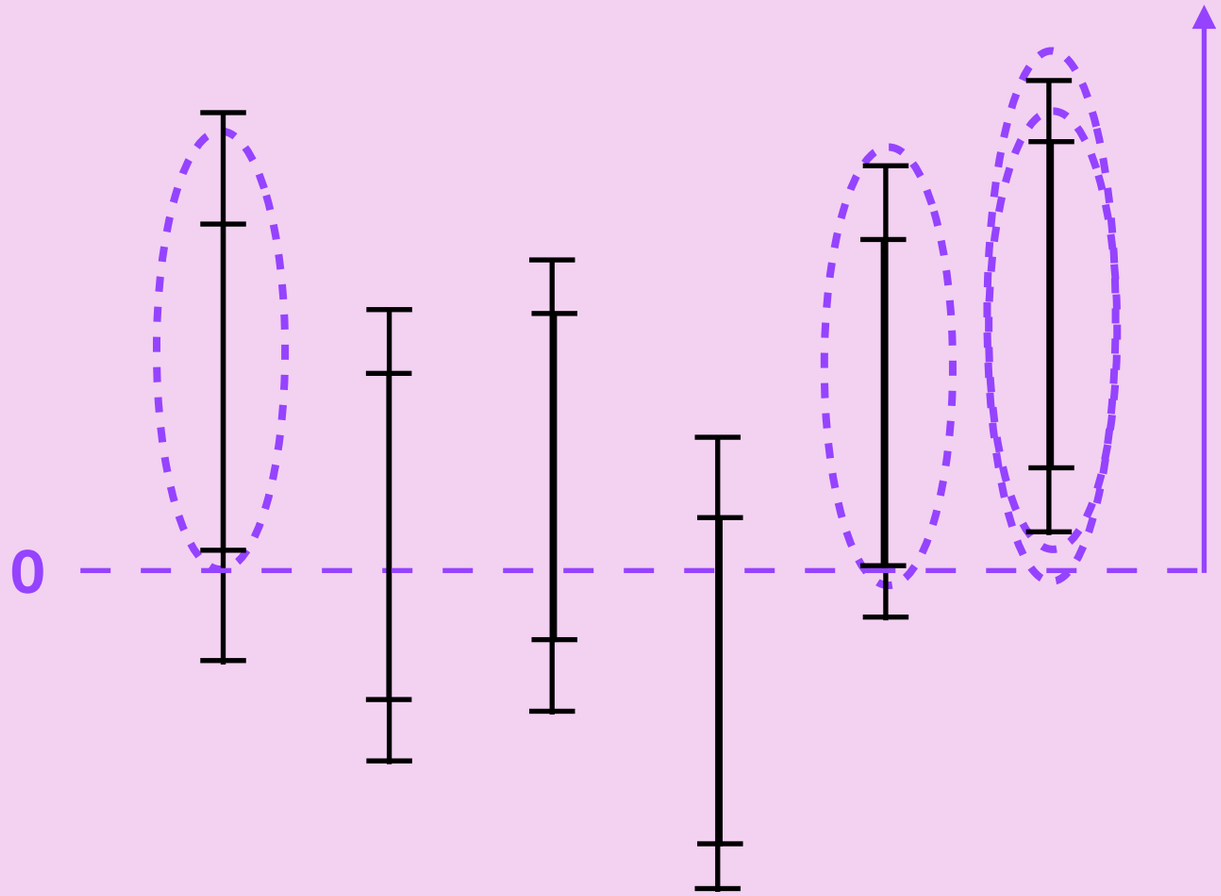
Analog of false discovery rate (FDR) from multiple testing

Benjamini and Yekutieli (2005) show how to
control FCR with corrected marginal CIs

# *Interlude*: for conditional coverage under arbitrary selection rules, we <u>can't</u> use marginal CIs

<u>Conditional coverage</u>: $\mathbb{P}(\theta_i^* \in C_i(\alpha_i) \mid i \in \mathcal{S}) \geq 1 - \alpha$

$\mathscr{R}$: select if 95% CI above 0
$\mathscr{R}$ (Bonferroni) : select if 99.2% CI above 0

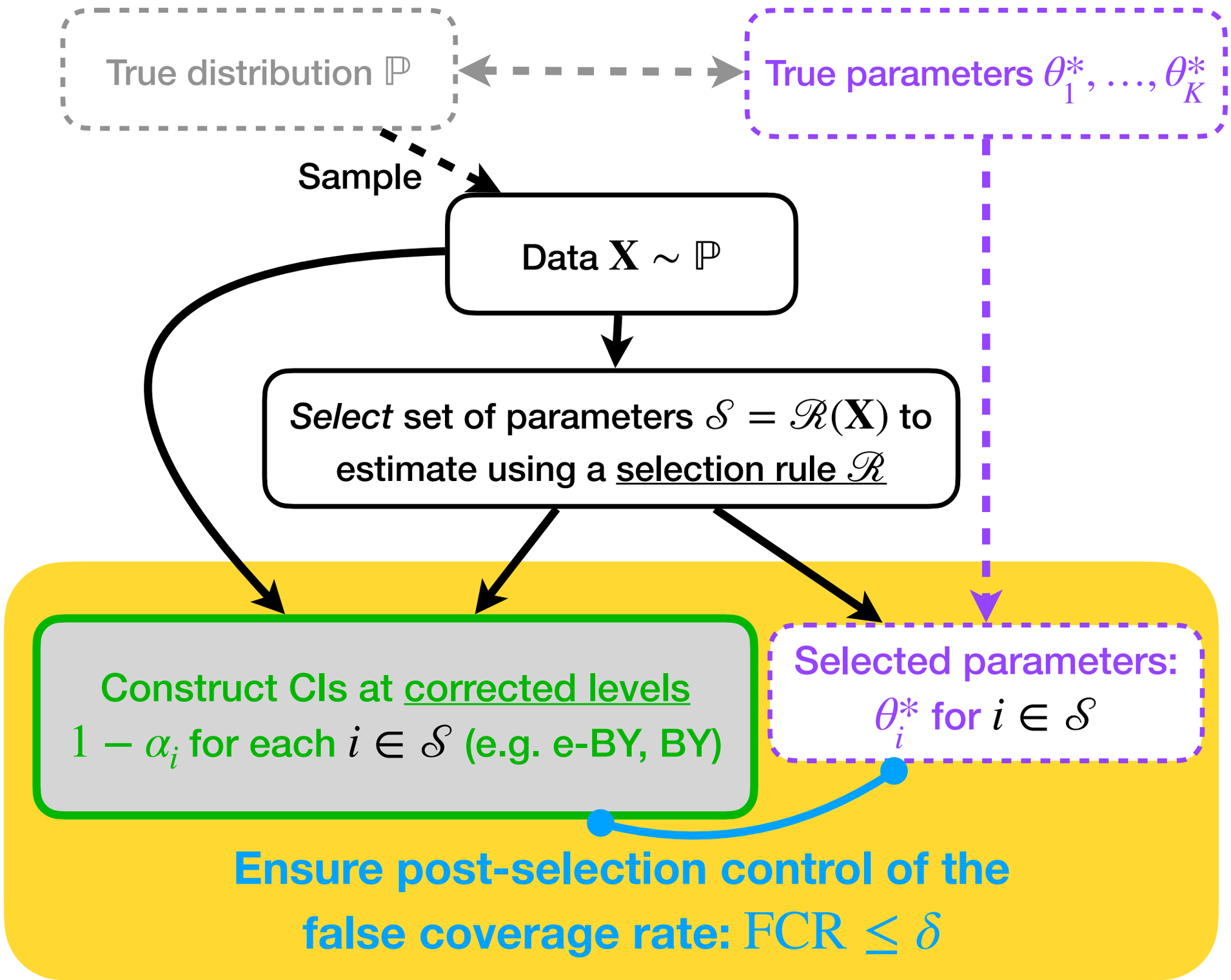



Assume we want 95% conditional coverage so we output $(1 - 0.05/K)$-CI for each $\theta_i^*$ (Bonferroni correction)

If the selection rule is "select $i$ when the $i$th $(1 - 0.05/K)$-CI is above the 0"…

Then, we *always* fail to cover when $\theta_i^* \leq 0$!

# Post-selection inference with FCR guarantees



# Outline

1. ~~Introduce the post-selection inference problem.~~

2. Compare our method (**e-BY procedure**) to the current state-of-the-art (**BY procedure**).

3. Describe a novel category of confidence intervals: **e-CIs.**

4. Results of simulations in a nonparametric setting.

# Current state-of-the-art: the BY procedure

We have access to marginal CIs for each $i \in \{1,\ldots,K\}$: $C_i(\alpha)$
s.t. $\mathbb{P}(\theta_i^* \in C_i(\alpha)) \geq 1-\alpha$ for any $\alpha \in (0,1)$.

**BY procedure:**

In the **known $\mathscr{R}$/independent (or PRDS)** case: output $C_i\left(\frac{\delta R_i^{\min}}{K}\right)$ for each $i \in \mathcal{S}$.

$1 \leq R_i^{\min} \leq |\mathcal{S}|$ is a value that depends on the selection rule.

In the **unknown $\mathscr{R}$/dependent** case: output $C_i\left(\frac{\delta|\mathcal{S}|}{K\ell_K}\right)$ for each $i \in \mathcal{S}$.

$\ell_K \approx \log K$ is the $K$th harmonic number.

**Theorem (BY 2005):** The BY procedure (above) ensures $\mathrm{FCR} \leq \delta$



Benjamini and Yekutieli (2005)

# Calculating $R_i^{\min}$ requires knowledge of the selection rule.

Recall that $\mathbf{X} = (X_1, \ldots, X_K)$ is our sampled data.

$R_i^{\min} := \min \{ |\mathscr{R}(X_1, \ldots, x_i, \ldots, X_K)| : x_i \in \mathscr{X}_i \text{ and } i \in \mathscr{R}(X_1, \ldots, x_i, \ldots, X_K)\}$

1. $X_i$ can be changed to any other possible data value $x_i$, but all other data $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k)$ remain fixed.
2. The $i$th parameter, $\theta_i$, remains in the resulting selection set with changed $x_i$.

Many *known* selection rules achieve the upper bound of $|\mathscr{S}|$ e.g. CI above threshold, Benjamini-Hochberg (BH) etc.

To compute $R_i^{\min}$, we require <u>full knowledge</u> of the selection rule $\mathscr{R}$.

Scientists arrive at selection sets in ad hoc ways, through messy analysis:
$\mathscr{R}$ may be impossible to describe!

$\mathscr{R}$ can be sophisticated even if well specified e.g. interactive, recursive procedure between CI construction and selection

**For unknown or ad-hoc selection rules,** no guarantees can be made i.e. cannot do better than $R_i^{\min} = 1$ (Bonferroni) and output $C_i\left(\frac{\delta}{K}\right)$.

Thus, fallback to $C_i\left(\frac{\delta |\mathscr{S}|}{K \ell_K}\right)$.

# Our method: the e-BY procedure

We have access to **an e-CI** ~~for each~~ $i \in \{1, \ldots, K\}$.

**e-BY procedure (ours):**

Output $C_i\left(\frac{\delta |\mathcal{S}|}{K}\right)$ for each $i \in \mathcal{S}$

**Theorem (ours):** The e-BY procedure above ensures $\mathrm{FCR} \leq \delta$

1. There is no reliance on selection rule (through $R_i^{\min}$) or change based on dependence structure.

2. E-BY requires access to e-CIs, a special class of CIs.

# Head-to-head comparison of e-BY vs. BY

| | e-BY | BY |
|---|---|---|
| **Knowledge of selection rule** | None required | Needed through $R_i^{\min}$ |
| **Correction factor** | Always output $C_i\left(\frac{\delta \lvert \mathcal{S} \rvert}{K}\right)$ | Know $\mathscr{R}$ and Independence/PRDS: $C_i\left(\frac{\delta R_i^{\min}}{K}\right)$<br>Otherwise: $C_i\left(\frac{\delta \lvert \mathcal{S} \rvert}{K \ell_K}\right)$ |
| **Type of CI** | ~~Only e-CIs~~<br>All CIs (calibrate each CI to an e-CI) | All CIs |

The BY procedure is a special case of the e-BY procedure obtained by calibrating CIs to e-CIs — **e-BY generalizes BY.**

# Outline

1. ~~Introduce the post-selection inference problem.~~

2. ~~Compare our method (**e-BY procedure**) to the current state-of-the-art (**BY procedure**).~~

3. Describe a novel category of confidence intervals: **e-CIs.**

4. Results of simulations in a nonparametric setting.

# E-value: e is for expectation (is bounded by 1)

$E(\theta)$ is an **e-value** w.r.t. a set of distributions with parameter $\theta$ if and only if:

1. $E(\theta)$ is nonnegative, and
2. $\mathbb{E}_\theta[E(\theta)] \leq 1$

E-values are analogs of p-values that have been extensively studied in recent work in testing and estimation (Shafer, Vovk, Grünwald, ourselves, and others).

**Fact:** $\mathbb{P}_\theta\left(E(\theta) \geq \frac{1}{\alpha}\right) \leq \alpha$ for any $\alpha \in (0,1)$.

True by Markov's inequality!

# e-CI: CI derived from an e-value

Denote the universe of parameters as $\Theta$.

$C$ is an **e-CI** if and only if there are e-values $E(\theta)$ where:

$$C(\alpha) := \left\{ \theta \in \Theta : E(\theta) < \frac{1}{\alpha} \right\}$$

**Fact:** Every e-CI $C$ is a valid confidence interval (CI).

**Proof:** Let $\theta^*$ be the true parameter.

$$\mathbb{P}_{\theta^*}(\theta^* \notin C(\alpha)) = \mathbb{P}_{\theta^*}\left( E(\theta^*) \geq \frac{1}{\alpha} \right) \leq \alpha$$

This is by Markov's inequality, again

# Proof of FCR control of e-BY

Recall e-BY outputs $C_i\left(\frac{\delta|\mathcal{S}|}{K}\right)$ for each $i \in \mathcal{S}$ where $C_i$ is an e-CI for $\theta_i$.

**Proof that $\mathrm{FCR} \leq \delta$ for e-BY:**

$$\mathbb{E}\left[\frac{\sum_{i\in\mathcal{S}} \mathbf{1}\left\{\theta_i^* \notin C_i\left(\frac{\delta|\mathcal{S}|}{K}\right)\right\}}{|\mathcal{S}| \vee 1}\right] = \mathbb{E}\left[\frac{\sum_{i\in\mathcal{S}} \mathbf{1}\left\{E_i(\theta_i^*) \geq \frac{K}{\delta|\mathcal{S}|}\right\}}{|\mathcal{S}| \vee 1}\right]$$

$$= \mathbb{E}\left[\frac{\sum_{i\in\mathcal{S}} \mathbf{1}\left\{E_i(\theta_i^*)\delta|\mathcal{S}|/K \geq 1\right\}}{|\mathcal{S}| \vee 1}\right] \leq \sum_{i=1}^{K} \mathbb{E}\left[\frac{E_i(\theta_i^*)\delta|\mathcal{S}|/K}{|\mathcal{S}| \vee 1}\right]$$

($\mathbf{1}\{x \geq 1\} \leq x$ and $\mathcal{S} \subseteq \{1,\ldots,K\}$)

$$\leq \frac{\delta}{K}\sum_{i=1}^{K} \mathbb{E}\left[E_i(\theta_i^*)\frac{|\mathcal{S}|}{|\mathcal{S}| \vee 1}\right] \leq \delta \quad \text{(def. of e-value)}$$

# Outline

1. ~~Introduce the post-selection inference problem.~~

2. ~~Compare our method (**e-BY procedure**) to the current state-of-the-art (**BY procedure**).~~

3. Describe a novel category of confidence intervals: **e-CIs.**

   A. Existing CIs are already e-CIs (universal inference, confidence sequences).

   B. CIs can be calibrated e-CIs (and BY is special case of e-BY)

4. Results of simulations in a nonparametric setting.

# Universal inference CI is an e-CI

Universal inference (Wasserman, Ramdas, Balakrishnan 2020) is a method for deriving e-CIs whenever the likelihood function is known.

Receive i.i.d. data $A_1, \ldots, A_n$ and split equally into two datasets $D_0, D_1$

For any $\theta \in \Theta$, perform a likelihood ratio test between:

$H_0 : \theta$ is the true parameter, $H_1 : \theta$ is not the true parameter

Estimate any likelihood $\hat{p}_1$ using $D_1$ (alternative likelihood)

$$C^{\mathrm{UI}}(\alpha) := \left\{ \theta \in \Theta : \alpha \hat{p}_1(D_0) < \max_{p \in P_\theta} p(D_0) \right\}.$$

$p$: maximum likelihood of $D_0$ under null (set of distributions with parameter $\theta$)

We can use universal inference to:

- estimate number of components in GMM in high dimensions
- estimate sparsity of regression problem
- determine if a distribution satisfy certain shape-constraints
- estimate parameters whenever we have likelihoods

# Stopped confidence sequences are e-CIs

In the sequential regime, samples come one at a time in a stream $A_1, A_2, \ldots$

An $(1-\alpha)$-**confidence sequence** is a sequence of intervals $(C^t(\alpha))_t$ where

$$\mathbb{P}(\forall t \in \mathbb{N} : \theta^* \in C^t(\alpha)) \geq 1-\alpha$$

or equivalently

$\mathbb{P}(\theta^* \in C^\tau(\alpha)) \geq 1-\alpha$ for any stopping time $\tau$ (i.e. data dependent time).

**Example (Howard et al. 2021):** If $A_i$ are 1-sub-Gaussian,

$$C^t(\alpha) := \frac{1}{t}\sum_{i=1}^{t} A_i \pm \sqrt{\frac{\log\log 2t + 0.72\log(10.4/\alpha)}{t}}$$

is a $(1-\alpha)$-confidence sequence for estimating $\theta^* = \mathbb{E}[A_i]$.

**Robust to stopping and sampling bias!**

**Theorem:** Stopped confidence sequences $(C^\tau(\alpha))$ are e-CIs.

**"Proof":** Confidence sequences are constructed by inverting nonnegative supermartingales.

Nonnegative supermartingales are e-values at stopping times.

We can build confidence sequences for any situation where we have Chernoff bounds (Howard et al. 2020, 2021).

We can also extend universal inference to the sequential regime.

Confidence sequences based off of nonnegative martingales are admissible (Ramdas et al. 2020).

# Calibration: deriving an e-CI from any CI

We can always *calibrate* a CI into an e-CI.

Based line of work about calibrating p-values into e-values (Shafer, Vovk, Wang, etc.).

A **calibrator** is a upper semicontinuous, nonincreasing function

$f : [0,1] \times [0,\infty]$ such that: $\int_0^1 f(x)\, dx \leq 1.$

Define $f^{-1}(x) = \sup\, \{p : f(p) \geq x\}.$

Let $C$ be an arbitrary CI.

**Theorem:** The following calibrated CI is an e-CI:

$$C^{\text{cal}}(\alpha) := C\left(f^{-1}\left(\frac{1}{\alpha}\right)\right).$$

Examples of calibrators:

- All or nothing: $f(p) = \frac{1}{\beta}\mathbf{1}\{p \leq \beta\}$ for any $\beta \in (0,1)$
- Power: $f(p) = \kappa p^{\kappa-1}$ for any $\kappa \in (0,1)$

# Calibration implies BY = e-BY under dependence

**The BY$(\delta, K)$ calibrator:**

$$f^{\mathrm{BY}(\delta,K)}(p) = \frac{K}{\delta} \cdot \frac{1}{\lceil K\ell_K p/\delta \rceil}$$

With CI $C_i$, recall that the BY procedure outputs for each $i \in \mathcal{S}$:

$$C_i\left(\frac{\delta\,|\,\mathcal{S}\,|}{K\ell_K}\right)$$

With the e-CI $C_i^{\mathrm{cal}}$calibrated with $f^{\mathrm{BY}(\delta,K)}$ from $C_i$, e-BY outputs:

$$C_i^{\mathrm{cal}}\left(\frac{\delta\,|\,\mathcal{S}\,|}{K}\right) = C_i^{\mathrm{cal}}\left(f^{\mathrm{BY}(\delta,K)^{-1}}\left(\frac{K\ell_K}{\delta\,|\,\mathcal{S}\,|}\right)\right) = C_i\left(\frac{\delta\,|\,\mathcal{S}\,|}{K\ell_K}\right)$$

The BY procedure is a special case of e-BY.

# Outline

1. ~~Introduce the post-selection inference problem.~~

2. ~~Compare our method (**e-BY procedure**) to the current state-of-the-art (**BY procedure**).~~

3. ~~Describe a novel category of confidence intervals: **e-CIs.**~~

   A. ~~Existing CIs are already e-CIs (universal inference, confidence sequences).~~

   B. ~~CIs can be calibrated e-CIs (and BY is special case of e-BY)~~

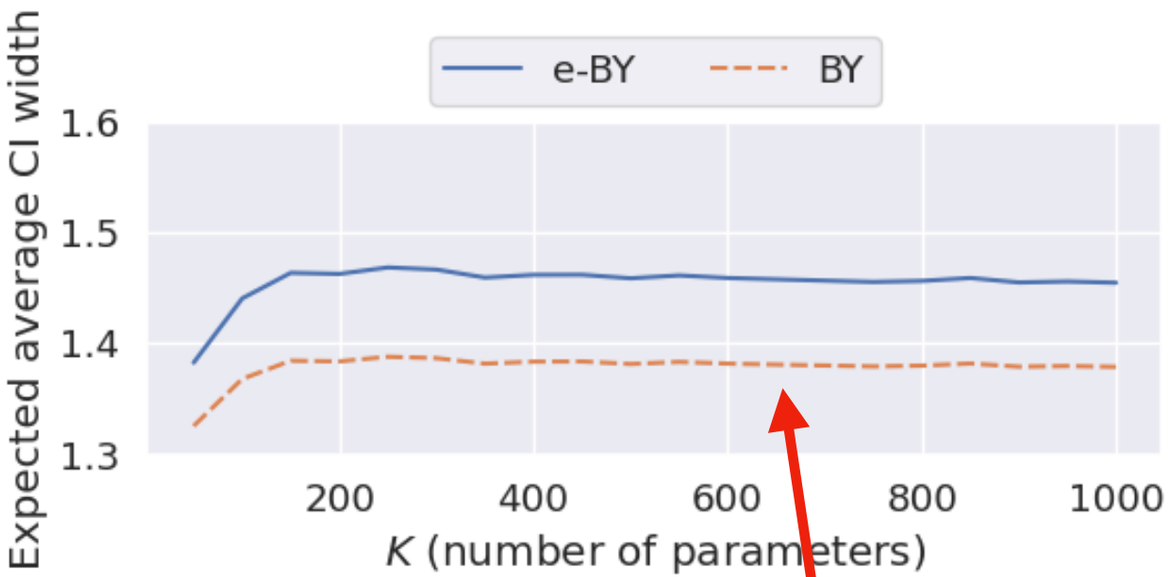
4. Results of simulations in a nonparametric setting.

# Simulations for bounded random variables indicate e-BY is tighter under dependence

Nonparametric setting: estimate the mean of bounded random variables in $[-1,1]$

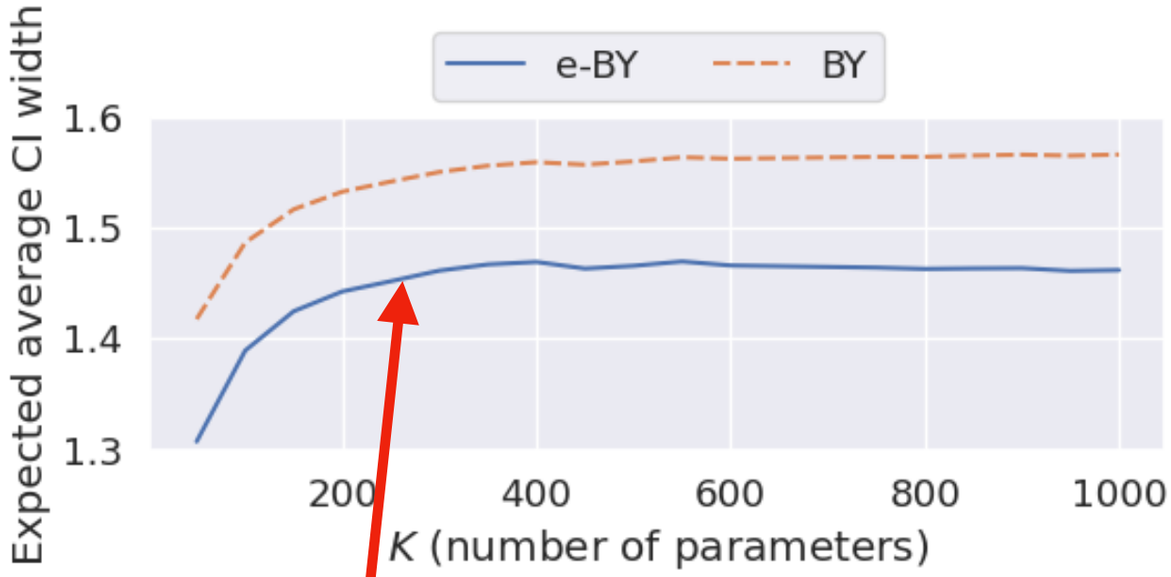
Hoeffding based CI for the BY procedure and e-CI for the e-BY procedure.

Select parameters that have solely positive $(1-\delta)$-CIs $\quad \delta = 0.1$




Known $\mathscr{R}$ and independent

BY tighter than e-BY




Unknown $\mathscr{R}$ or dependent

e-BY tighter than BY

# Takeaways

1. e-BY procedure provides FCR control with no assumptions about dependence and the selection rule, as opposed to the BY procedure.

   A. Use BY only when selection rule is known *and* the data is independent.

   B. Otherwise (unknown selection rule or dependent data), use e-BY.

2. BY is a special case of e-BY.

3. e-CIs can are already used in many settings e.g. universal inference, sequential settings, Chernoff methods.

4. e-CIs are particularly tight in the sequential regime — **this makes e-BY robust to stopping, sampling, and selection bias**.

5. **(In paper)** e-BY has sharp FCR control, and is an admissible inference procedure.

**Thanks!** **arXiv: 2203.12572**