Statistical Inference for Adaptive Experimentation

Kelly W. Zhang

Harvard John A. Paulson School of Engineering and Applied Sciences

Thesis Defense. April 26, 2023.

Advisors: Susan Murphy and Lucas Janson

NSF NIH Siebel Scholars

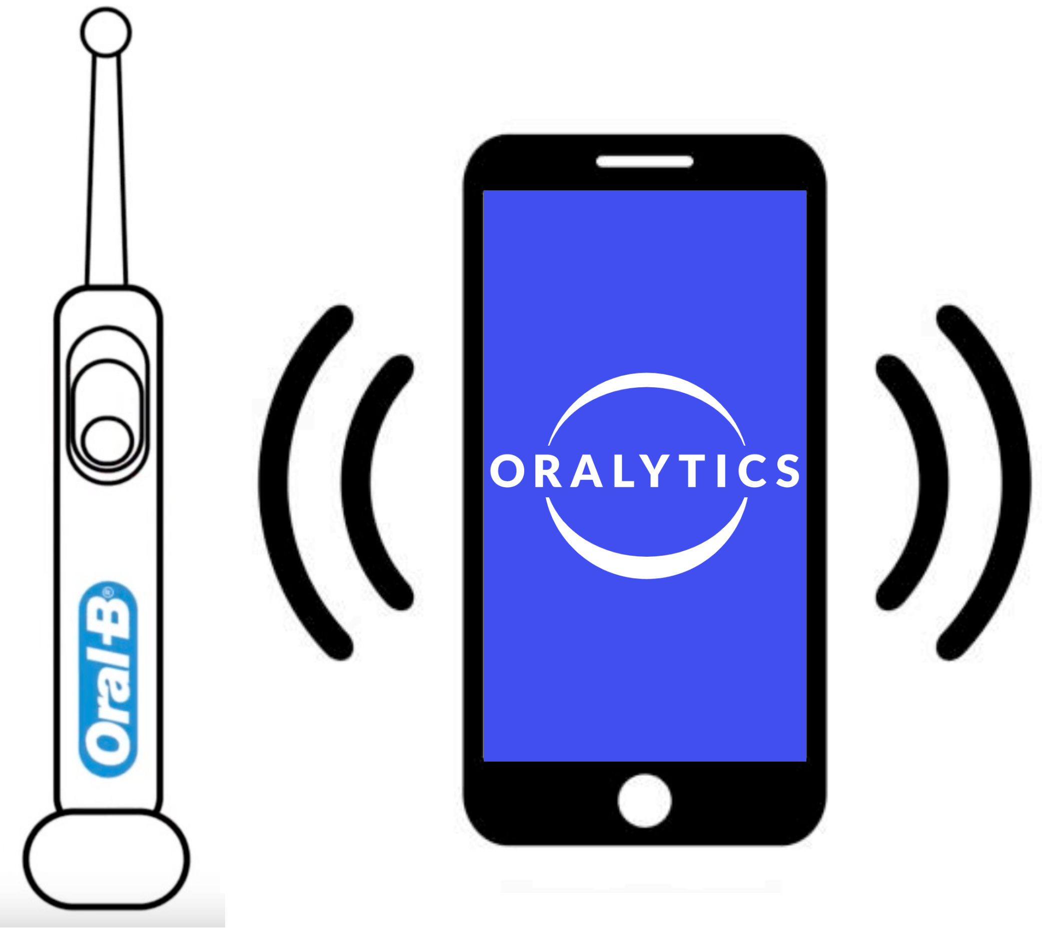
Our lives are becoming increasingly digitalized…

Healthcare

Public Policy

Education

Opportunity: Develop Digital Interventions



UNIVERSITY OF MICHIGAN

UCLA School of Dentistry

Digital Oral Health Coaching

U.S. Department of Veterans Affairs

PTSD Coach

PTSD Family Coach

Stay Quit Coach

Insomnia Coach

Mobile Health Apps Developed by U.S. Veterans Affairs

Apple HealthAI

Challenge: Learning what interventions to deliver—and when

Minimize: User Burden

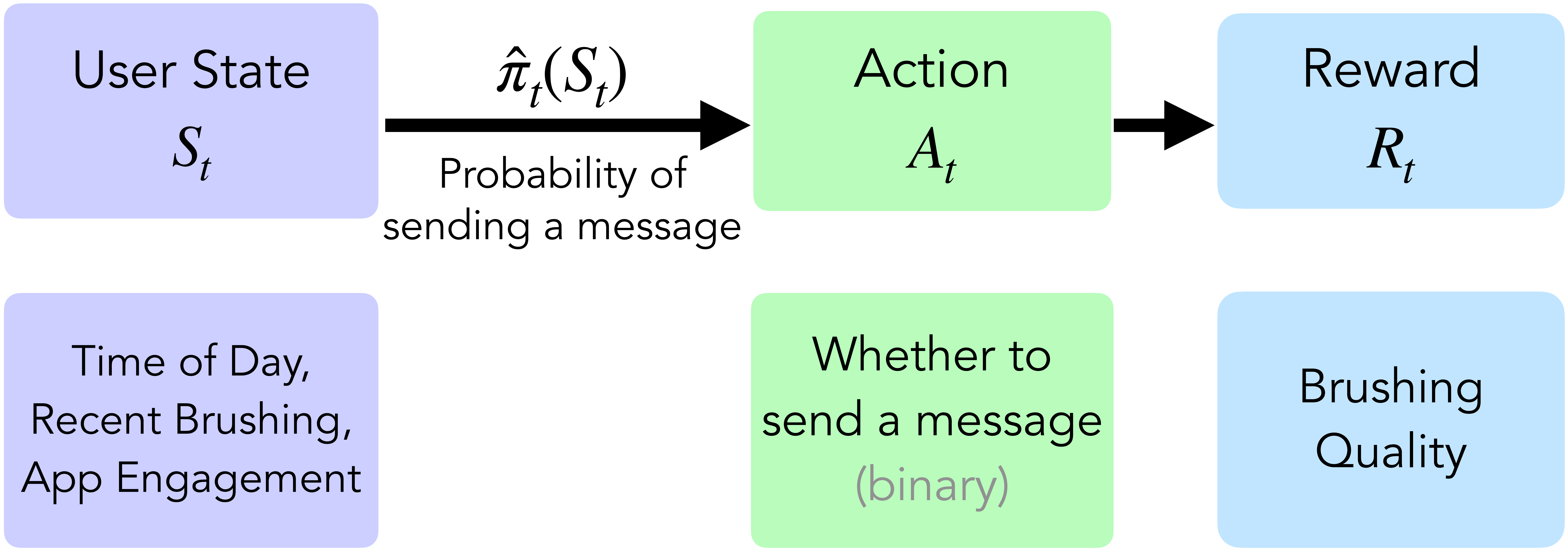
Maximize: User Benefit

Challenge: Learning what interventions to deliver—and when

Minimize:
User Burden

Maximize:
User Benefit

Online Reinforcement Learning (RL)



S_t, A_t, R_t definitions are design decisions

Oralytics Setting

Use (S_t, A_t, R_t) to update and form $\hat{\pi}_{t+1}$

My research focus is developing methodology to facilitate real-world deployments of online RL for digital interventions

Causal Inference for Sequential Decision Making

Designing Practical RL Algorithms for Real-World Deployments

Digital Intervention Study Design Objectives

Within-Study Personalization

Maximize User Benefit

- Send messages at opportune moments

Use Online RL Algorithms to maximize $\mathbb{E}\left[\sum_{t=1}^{T} R_t\right]$

After-Study Analyses

Evaluate the Intervention

- Understand heterogeneity across user types and user states

Infer Treatment Effects

$$\mathbb{E}\left[R_t \mid S_t, A_t = 1\right] - \mathbb{E}\left[R_t \mid S_t, A_t = 0\right]$$

Digital Intervention Study Design Objectives

Within-Study Personalization

Maximize User Benefit

- Send messages at opportune moments

Use Online RL Algorithms to maximize $\mathbb{E}\left[\sum_{t=1}^{T} R_t\right]$

After-Study Analyses

Confidence Intervals Critical for

- Replicable science
- Publishing and sharing results

Infer Treatment Effects

$$\mathbb{E}\left[R_t \mid S_t, A_t = 1\right] - \mathbb{E}\left[R_t \mid S_t, A_t = 0\right]$$

RL Algorithms Induce Dependence

Data tuples $\left(S_t, A_t, R_t\right)$ **are not independent over** $t \in [1: T]$

- RL data is "adaptively collected"

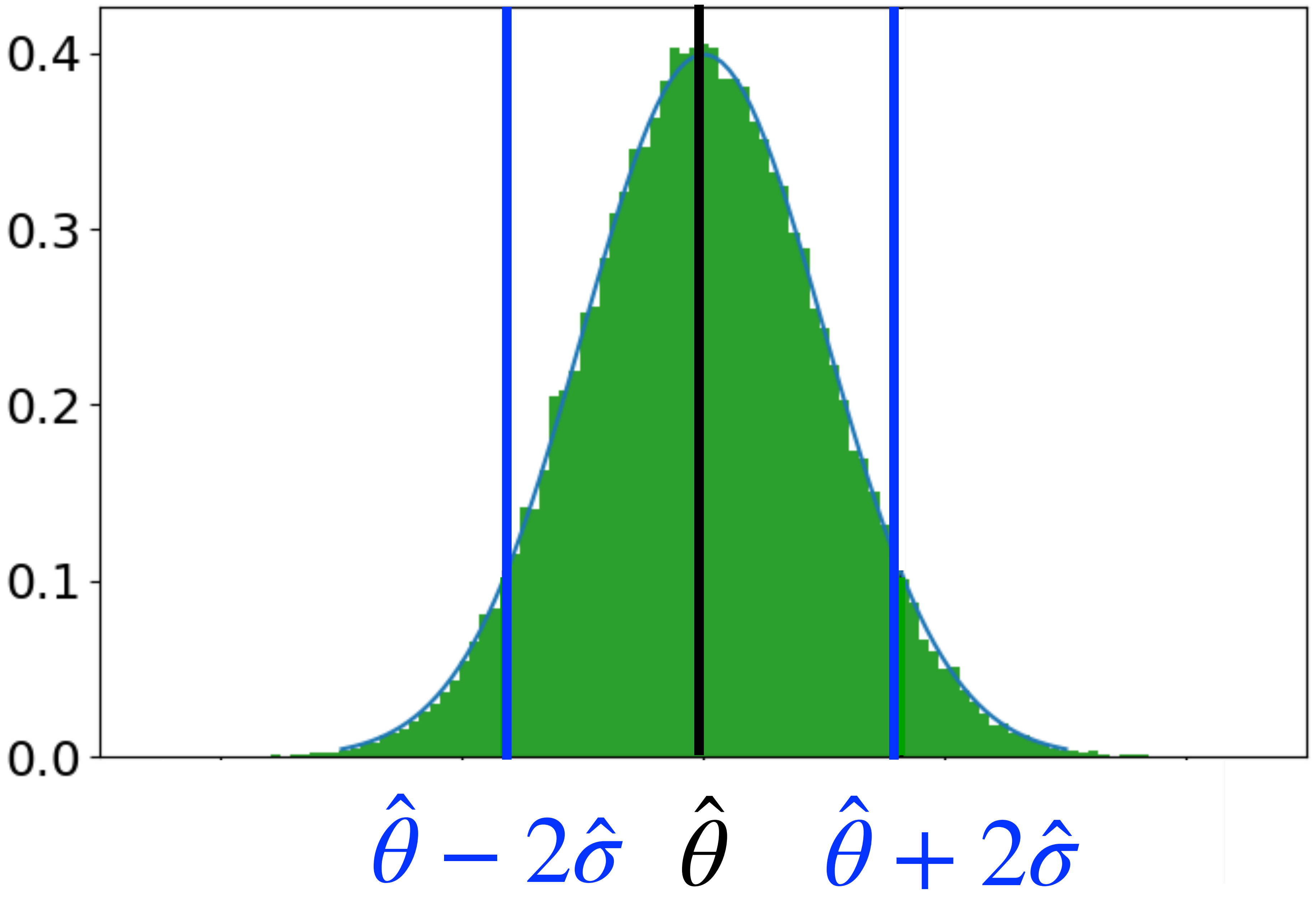
Consequences for Statistical Inference

- Bias [Nie et al.,'18] [Shin, Ramdas, Rinaldo; '19, '20]
- Asymptotic Non-Normality **[Zhang, Janson, Murphy; '20]**

Consequences of Dependence for Statistical Inference

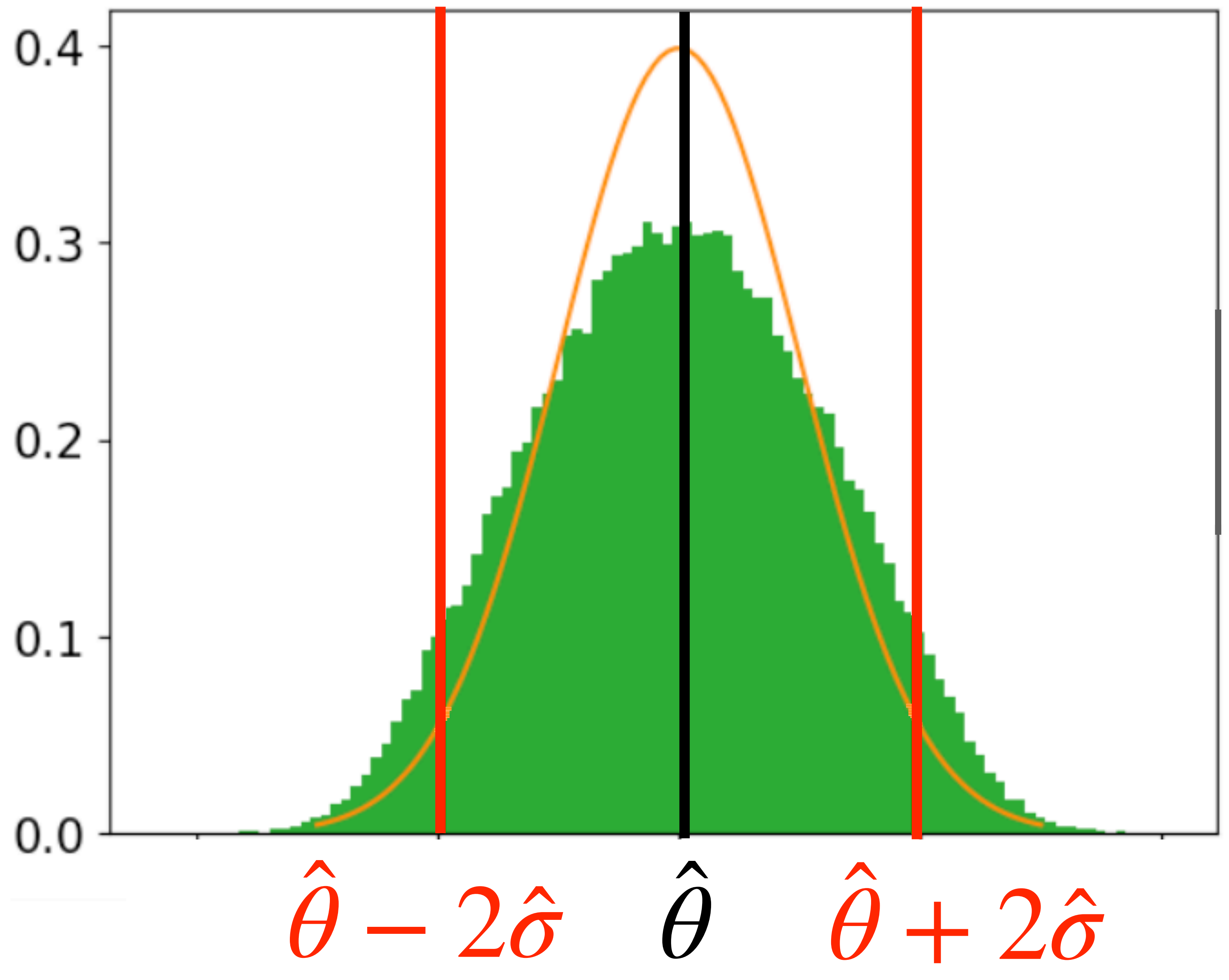
[**Zhang**, Janson, & Murphy, NeurIPS 2020]

Difference in Sample Means
Independently Collected Data



95% Percent Confidence Interval

Difference in Sample Means
Under Thompson Sampling



Only 89.5% coverage (expect 95%)

Statistical Inference after Using Online RL

Contributions

Inference for Batched Bandits
NeurIPS 2020
Zhang, Janson, & Murphy

Statistical Inference for M-Estimators on Adaptively Collected Data
NeurIPS 2021
Zhang, Janson, & Murphy

Statistical Inference Adaptive Sampling for Longitudinal Data
Under review
Zhang, Janson, & Murphy

Statistical Inference after Using Online RL

Contributions

Inference for Batched Bandits
NeurIPS 2020
Zhang, Janson, & Murphy

Statistical Inference for M-Estimators on Adaptively Collected Data
NeurIPS 2021
Zhang, Janson, & Murphy

Statistical Inference Adaptive Sampling for Longitudinal Data
Under review at Annals of Statistics
Zhang, Janson, & Murphy

Impact / Use Cases

Political Science: Survey Methods to Understand Voter Views

Offer-Westort, Coppock, & Green, 2022

THE UNIVERSITY OF CHICAGO Yale COLUMBIA UNIVERSITY

Statistical Inference after Using Online RL

Contributions

Inference for Batched Bandits
NeurIPS 2020
Zhang, Janson, & Murphy

Statistical Inference for M-Estimators on Adaptively Collected Data
NeurIPS 2021
Zhang, Janson, & Murphy

Statistical Inference Adaptive Sampling for Longitudinal Data
Under review
Zhang, Janson, & Murphy

Impact / Use Cases

Education: Automated Phone Calls to Encourage Parental Involvement

Esposito & Sautmann, 2022

WORLD BANK GROUP

Statistical Inference after Using Online RL

Contributions

Inference for Batched Bandits
NeurIPS 2020
Zhang, Janson, & Murphy

Statistical Inference for M-Estimators on Adaptively Collected Data
NeurIPS 2021
Zhang, Janson, & Murphy

Statistical Inference Adaptive Sampling for Longitudinal Data
Under review
Zhang, Janson, & Murphy

Impact / Use Cases

Digital Health: Enables use of online RL algorithms that combine data across users to learn

Oralytics

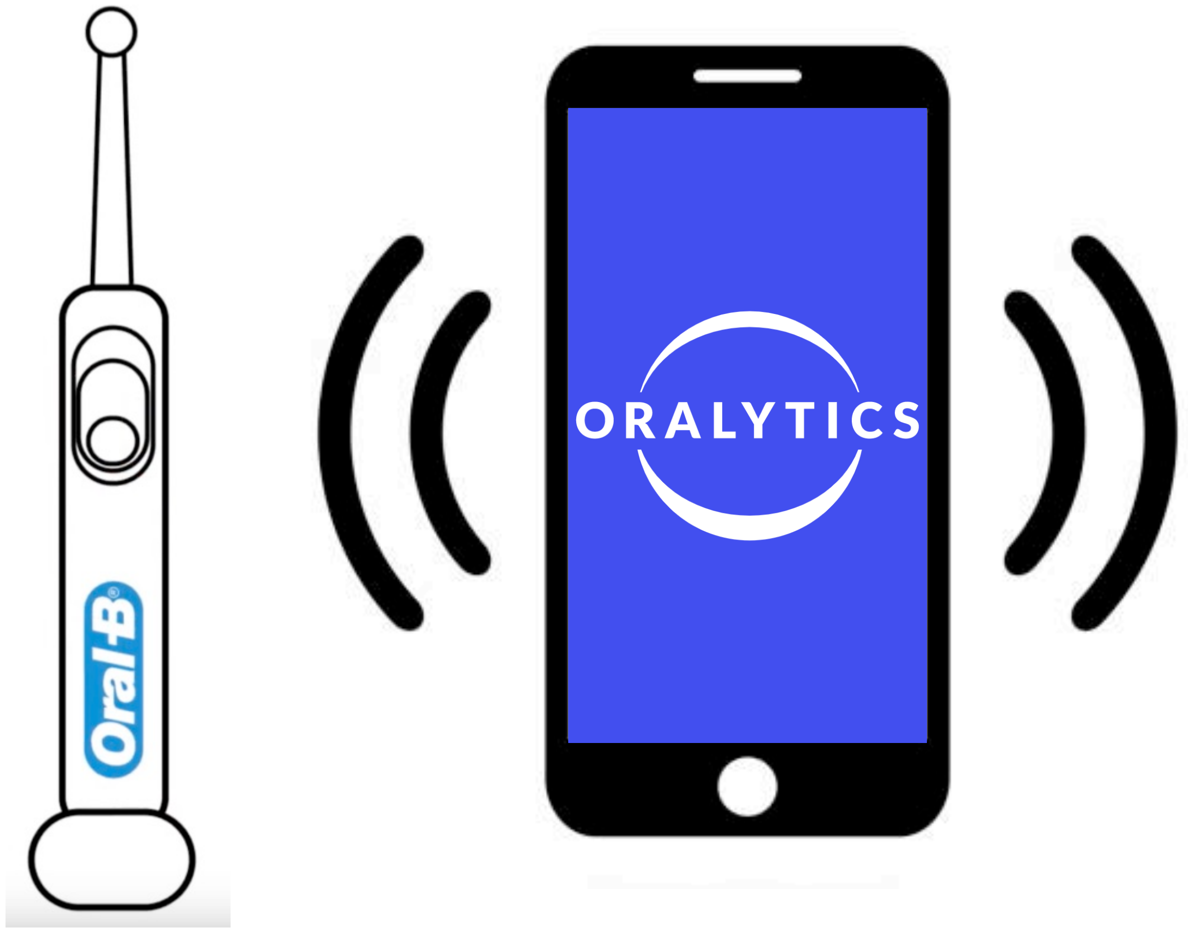
MiWaves

Talk Overview

Part 1:
Contextual Bandit Setting

Online Advertising

Part 2:
Longitudinal Data Setting



Digital Health

Part 1: Contextual Bandit Environment

Online Advertising Setting

At each decision time

$t \in [1:T]$ we see a new user

User State S_t → Action A_t → Reward R_t

Demographic Info

Type of ad

Sign up outcome

Contextual Bandit Environment

Potential Outcomes:
$\{S_t,\ R_t(0),\ R_t(1)\}_{t=1}^{T}$ i.i.d. over t

Data Tuple: $D_t = \left(S_t, A_t, R_t\right)$

Action selection probabilities:
$\mathbb{P}\left(A_t = 1 \,|\, D_{1:t-1}, S_t\right)$

(S_t, A_t, R_t) dependent over time $t \in [1 : T]$!!

Potential Outcomes	$t = 1$	$t = 2$	$t = T$	...	$t = T$
States	S_1	S_2	S_3	...	S_T
Rewards Under Action 0	$R_1(0)$	$R_2(0)$	$R_3(0)$	...	$R_T(0)$
Rewards Under Action 1	$R_1(1)$	$R_2(1)$	$R_3(1)$	...	$R_T(1)$
Actions Selected by RL Algorithm	$A_1 = 0$	$A_2 = 1$	$A_3 = 1$	...	$A_T = 0$

Blue indicates observed data

Inferential Goal

Parameters in an outcome model

- **Linear Model:** $\mathbb{E}\left[R_t \mid S_t, A_t\right] = S_t^\top \theta_0^\star + A_t S_t^\top \theta_1^\star$
- **Logistic Model:** $\mathbb{E}\left[R_t \mid S_t, A_t\right] = \left[1 + \exp\left(S_t^\top \theta_0^\star + A_t S_t^\top \theta_1^\star\right)\right]^{-1}$
- **Poisson Model:** $\mathbb{E}\left[R_t \mid S_t, A_t\right] = \log\left[S_t^\top \theta_0^\star + A_t S_t^\top \theta_1^\star\right]$

Interested in Treatment Effect

$$\mathbb{E}\left[R_t \mid S_t, A_t = 1\right] - \mathbb{E}\left[R_t \mid S_t, A_t = 0\right]$$

Inferential Goal

Treatment effect parameter $\theta_1^\star$

Parameters in an outcome model

- **Linear Model:** $\mathbb{E}\left[R_t \mid S_t, A_t\right] = S_t^\top \theta_0^\star + A_t S_t^\top \theta_1^\star$
- **Logistic Model:** $\mathbb{E}\left[R_t \mid S_t, A_t\right] = \left[1 + \exp\left(S_t^\top \theta_0^\star + A_t S_t^\top \theta_1^\star\right)\right]^{-1}$
- **Poisson Model:** $\mathbb{E}\left[R_t \mid S_t, A_t\right] = \log\left[S_t^\top \theta_0^\star + A_t S_t^\top \theta_1^\star\right]$

Interested in Treatment Effect

$$\mathbb{E}\left[R_t \mid S_t, A_t = 1\right] - \mathbb{E}\left[R_t \mid S_t, A_t = 0\right]$$

Typical Approach to Forming Estimators

Estimator $\hat{\theta}$ minimizes empirical loss:

$$\hat{\theta} \triangleq \operatorname{argmin} \frac{1}{T} \sum_{t=1}^{T} \ell_{\theta}(R_t, S_t, A_t)$$

Examples

- Sample mean
- Least squares
- Logistic regression
- Maximum likelihood

Typical Approach to Forming Estimators

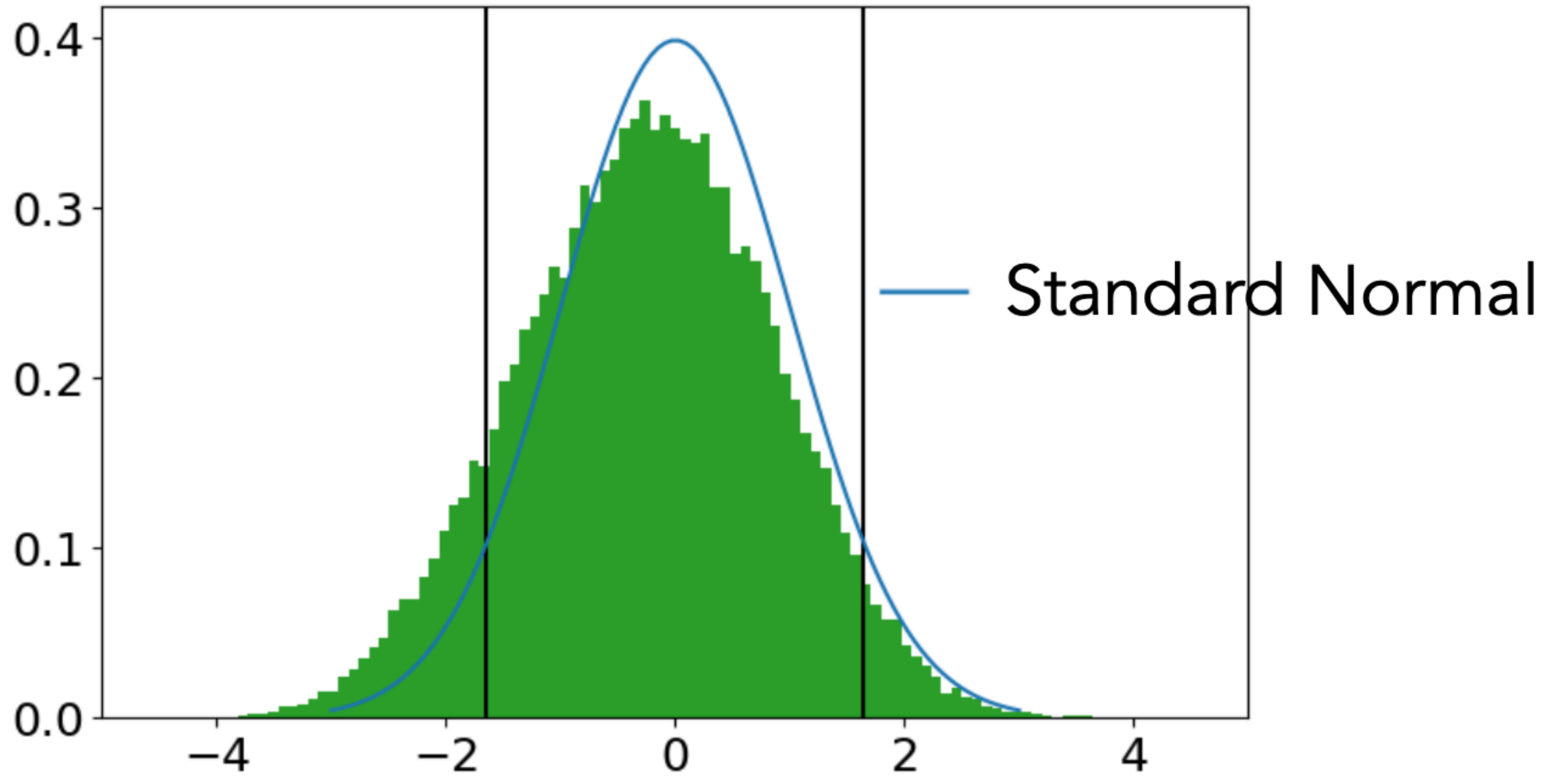
Estimator $\hat{\theta}$ minimizes empirical loss:

$$\hat{\theta} \triangleq \arg\min \frac{1}{T} \sum_{t=1}^{T} \ell_\theta(R_t, S_t, A_t)$$

Examples

- Sample mean
- Least squares
- Logistic regression
- Maximum likelihood

Empirical Distribution of Z-Statistic for the Sample Mean



Coverage: 84.9%
(Nominal 90%)

Thompson Sampling;
$\mathcal{N}(0,1)$ errors; $T = 1000$

Previous Approaches

Inference after Adaptive Sampling

[Hadad et al., 2021; Bibaut et al. 2021; Zhan et al. 2022; Deshpande et al., 2018]

- Off policy evaluation and infer parameters in simple models
- Cannot be used to infer parameters of general models

High Probability Bounds

[Abbasi-Yadkori et al., 2011; Kaufman et al., 2018; Jamieson et al., 2014; Howard et al., 2021]

- Finite sample guarantees
- Conservative - need much larger sample sizes

Adaptive Weighting Approach

Estimator $\hat{\theta}$ minimizes empirical loss:

$$\hat{\theta} \triangleq \operatorname{argmin} \frac{1}{T} \sum_{t=1}^{T} W_t \, \ell_\theta(R_t, S_t, A_t)$$

Adaptive *Square-Root* Inverse Propensity Weights

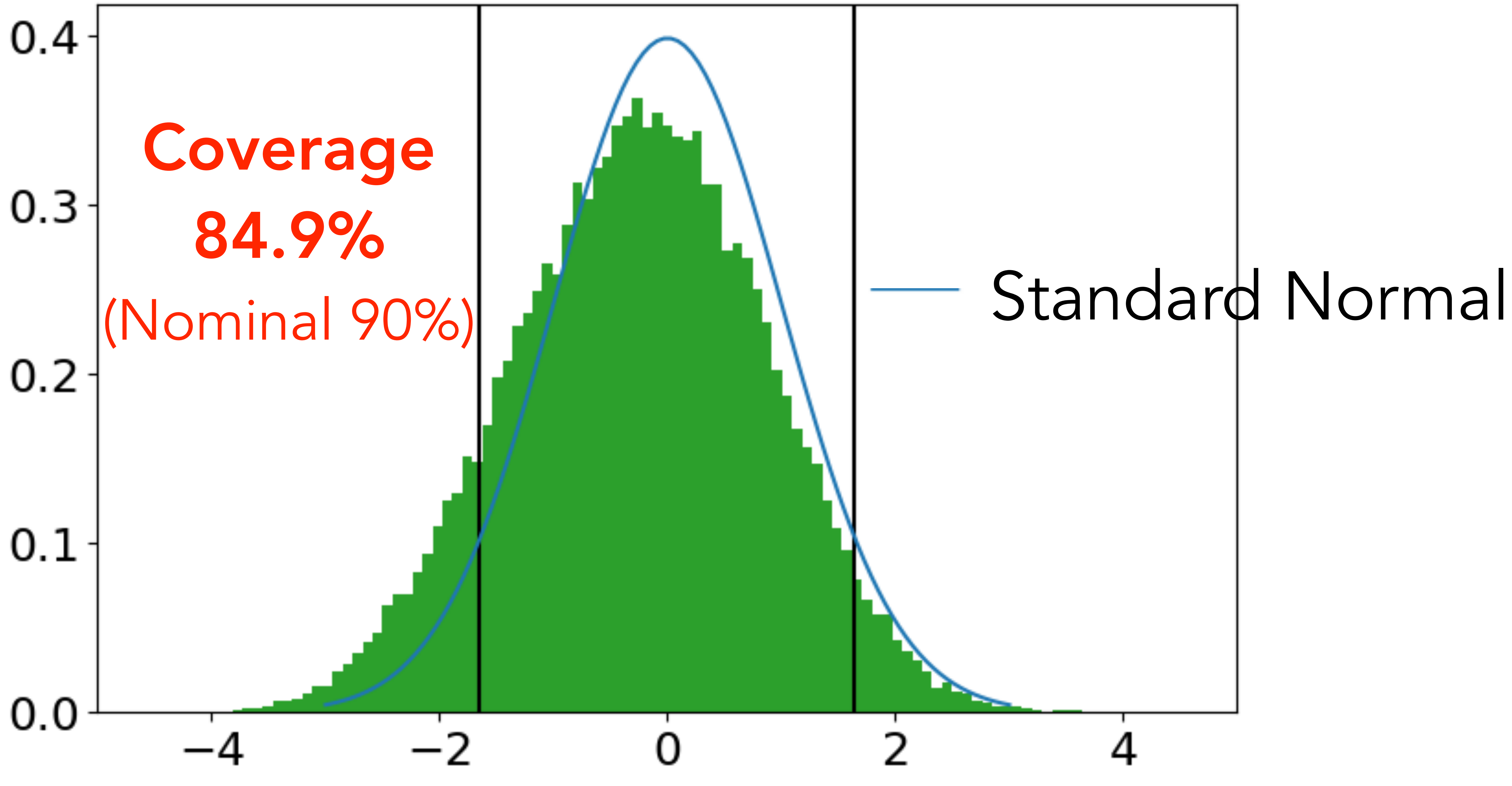
$$W_t = \frac{1}{\sqrt{\mathbb{P}(A_t \mid D_{1:t-1}, S_t)}}$$

Examples

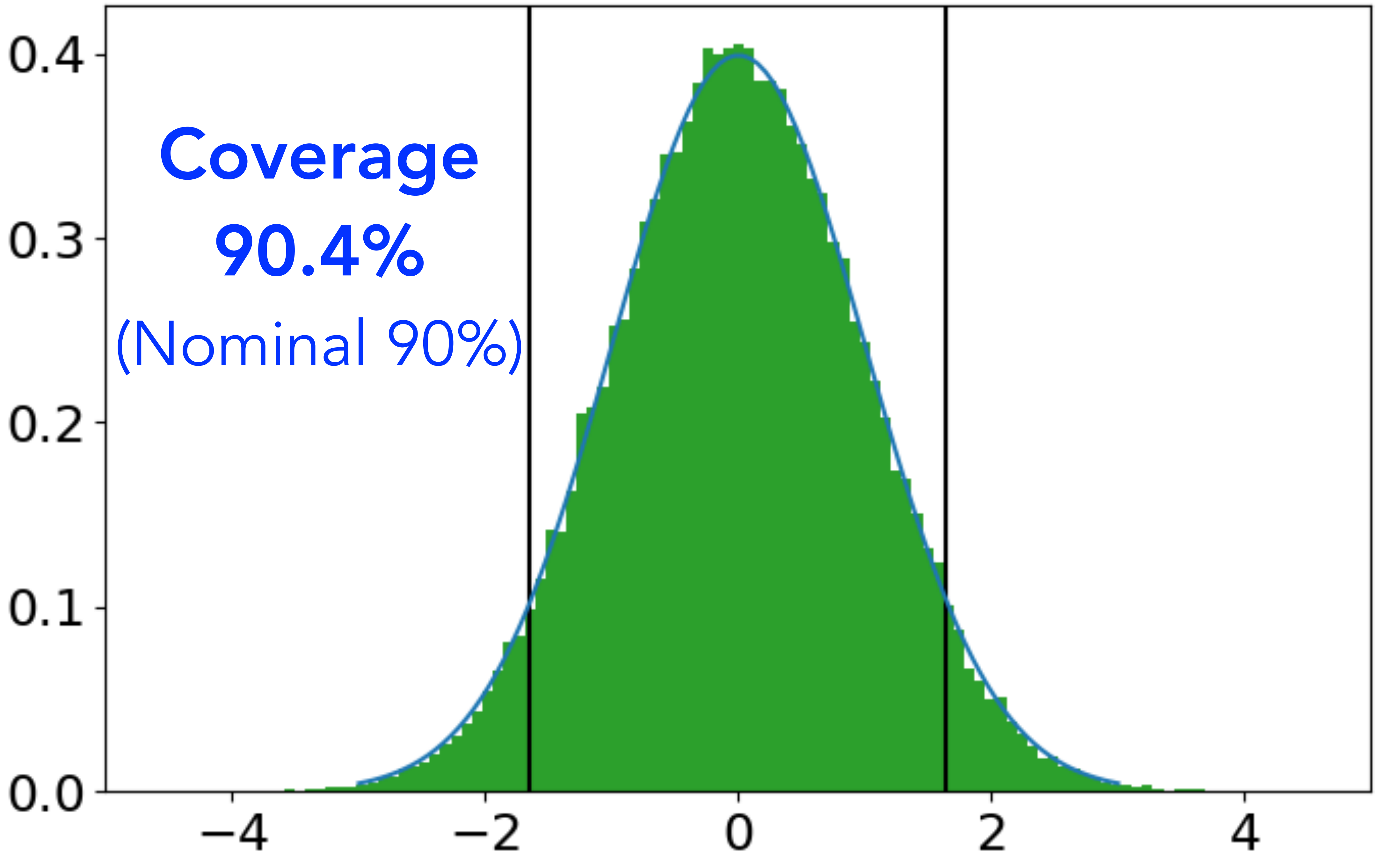
- Weighted least squares
- Weighted logistic regression
- Weighted maximum likelihood

Our Solution: Include "Adaptive" Weights

Empirical Distribution of Z-Statistic
(Unweighted) Sample Mean



Empirical Distribution of Z-Statistic
Adaptively Weighted Sample Mean



- Two-arm bandit with $T = 1000$
- **Thompson Sampling** with standard normal priors

Asymptotic Normality Result with Adaptive Weighting

$\theta^\star$ satisfies $\theta^\star \triangleq \operatorname{argmin} \mathbb{E}\left[\ell_\theta(R_t, S_t, A_t) \middle| S_t, A_t\right]$ for all S_t, A_t

$$\left\{ \frac{1}{T} \sum_{t=1}^{T} W_t \, \ddot{\ell}_{\hat{\theta}} \left(R_t, S_t, A_t\right) \right\} \sqrt{T} \left(\hat{\theta} - \theta^\star\right) \xrightarrow{D} N(0, \Sigma)$$

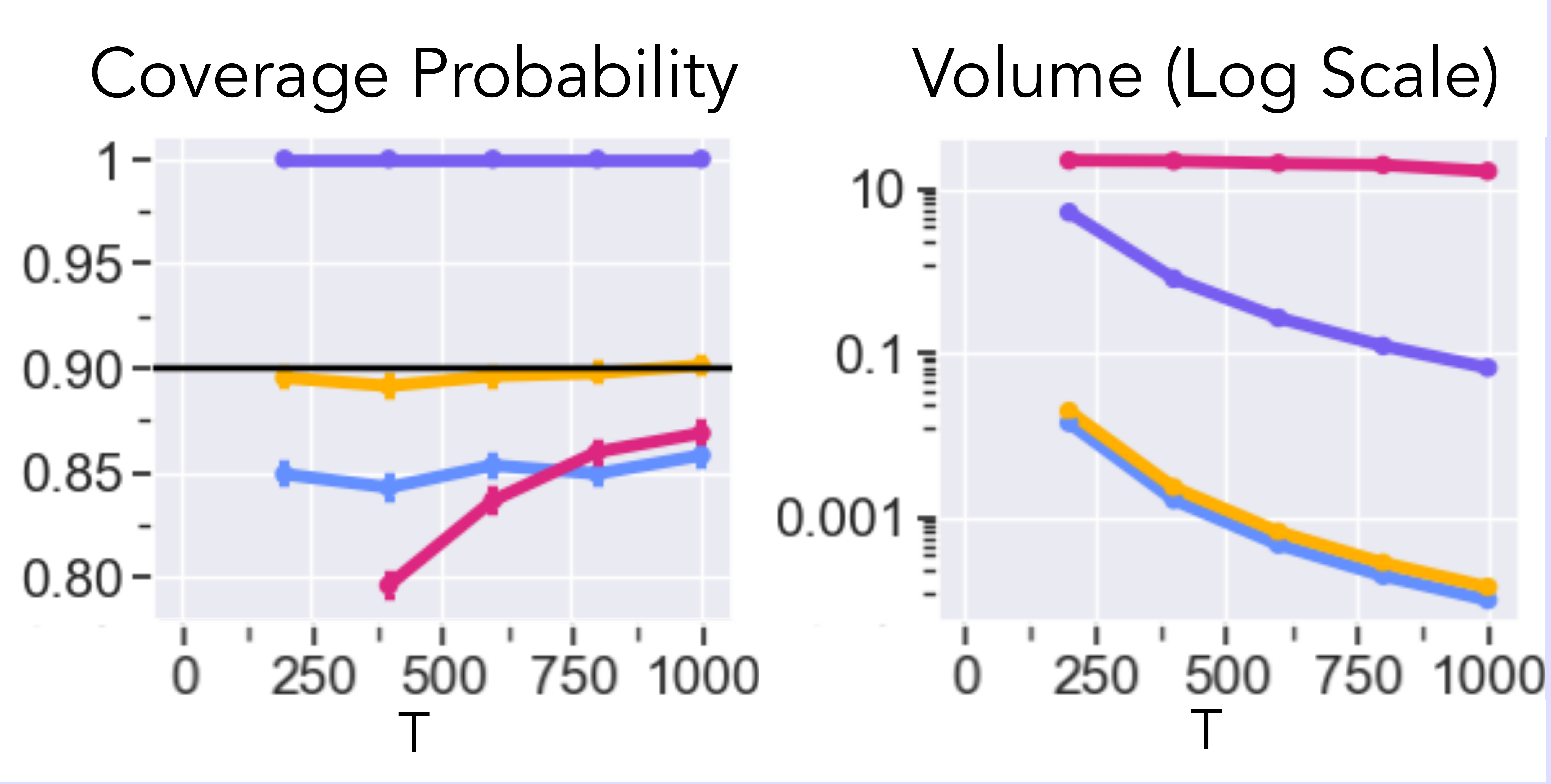
$$\Sigma = \mathbb{E}\left[\dot{\ell}_\theta \left(R_t, S_t, A_t\right) \left\{ \dot{\ell}_\theta \left(R_t, S_t, A_t\right) \right\}^\top \right]$$

Weighted Least Squares

Confidence Regions for $\theta^\star = [\theta_0^\star, \theta_1^\star]$ where

$$\mathbb{E}\left[R_t \mid A_t, S_t\right] = S_t^\top \theta_0^\star + A_t S_t^\top \theta_1^\star$$

90% Confidence Regions



- Least Squares (unweighted)
- W-Decorrelated [Deshpande et al., 2018]
- Self-Normalized Martingale Bound [Abbasi-Yadkori et al., 2011]
- Adaptively Weighted Least Squares

Similar performance for generalized linear models for Bernoulli and Poisson rewards

Role of adaptive weights

$$\hat{\theta} \triangleq \operatorname{argmin} \frac{1}{T} \sum_{t=1}^{T} W_t \, \ell_\theta(R_t, S_t, A_t)$$

$$W_t = \frac{1}{\sqrt{\mathbb{P}(A_t \mid D_{1:t-1}, S_t)}}$$

Adaptive weights are ***not*** used for

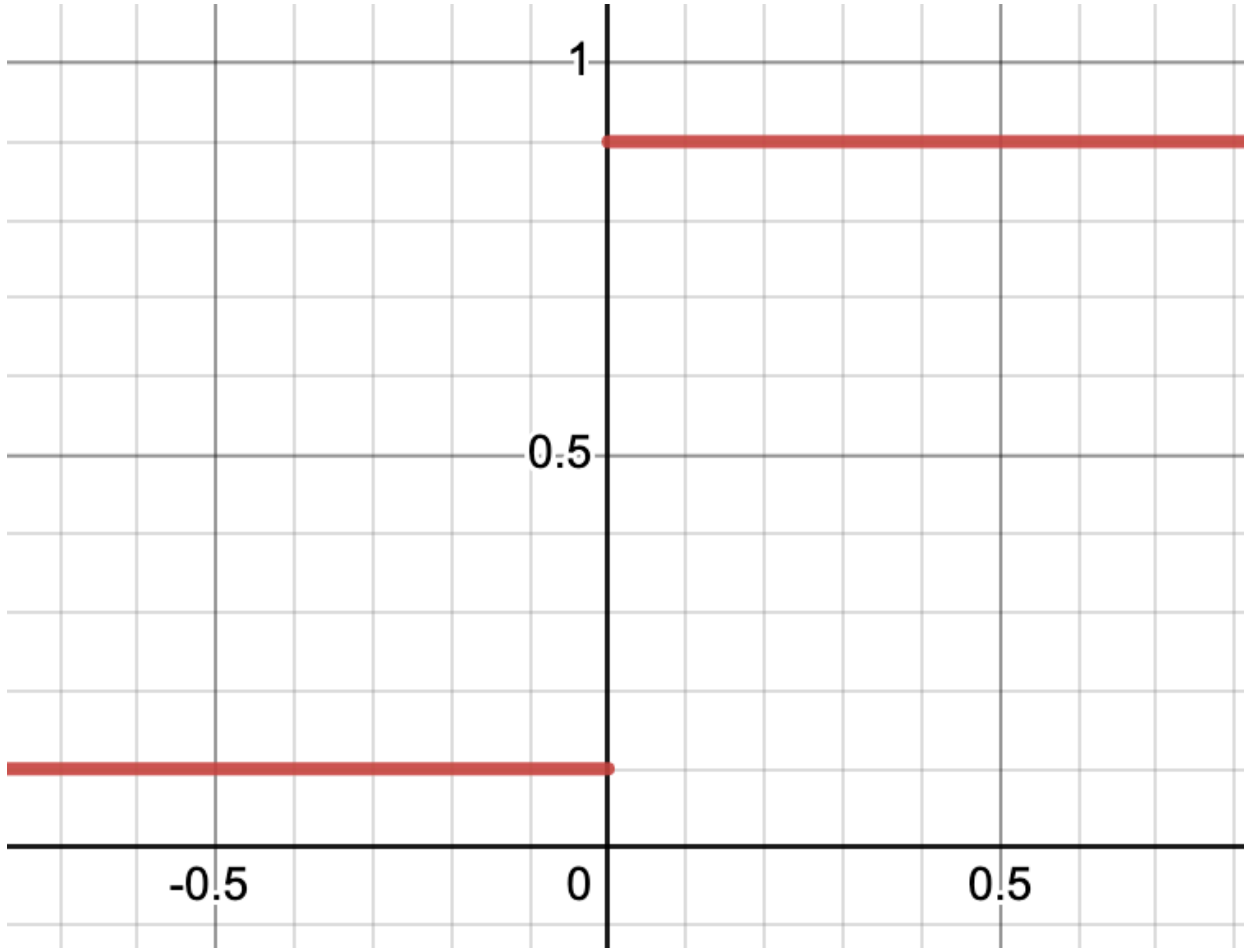
- Adjusting for heteroskedastic errors
- Defining the estimand (e.g. in causal inference, off-policy evaluation)

Used to **"stabilize" the variance** of the estimator due to instability of the adaptive policy

Instability of the Adaptive Policy

Limiting Action Selection Probabilities

Probability of Selecting $A_t = 1$



Treatment Effect:

$$\mathbb{E}\left[R_t(1)\right] - \mathbb{E}\left[R_t(0)\right]$$

Other examples non-smoothness problems:

- CI for test error of classifier
- Bootstrap
- Hodges estimator

Summary

- Common RL algorithms can form policies that are unstable
- Including adaptive weights
  - "Stabilizes" the variance of estimators
  - Ensures asymptotic normality
- Limitation
  - Approach not applicable to longitudinal data settings (multiple decision times per user)

Part 2: Longitudinal Data Setting

Oralytics Setting

Make a series of decisions for each user $i \in [1:N]$

User State $S_{i,t}$ → **Action** $A_{i,t}$ → **Reward** $R_{i,t}$

User State $S_{i,t}$	Action $A_{i,t}$	Reward $R_{i,t}$
Time of day, Previous brushing, App engagement	Whether to send message	Brushing quality

Oralytics Study Overview

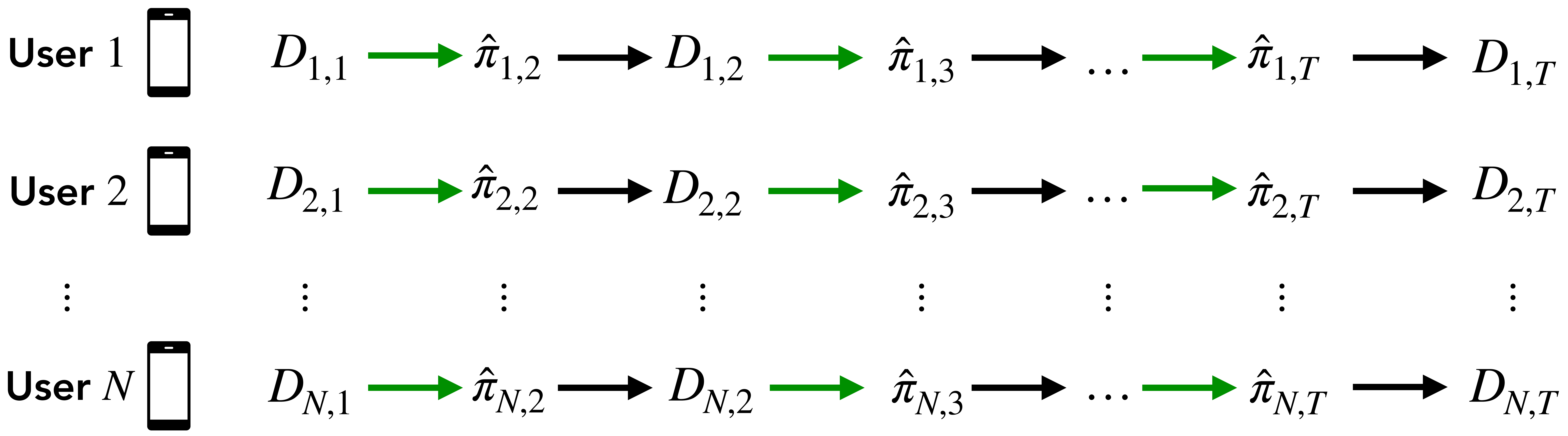
- **Total Decision Times:** 10 weeks with two decision times per day ($T = 140 = 10 \cdot 7 \cdot 2$)
- **Study Population:** $N \approx 70$ patients from dental clinics in Los Angeles
- **Data Collected After Study:** For each user $i \in [1 : N]$,

$$\underbrace{\left(S_{i,1}, A_{i,1}, R_{i,1}\right)}_{D_{i,1}} \quad \underbrace{\left(S_{i,2}, A_{i,2}, R_{i,2}\right)}_{D_{i,2}} \quad \cdots \quad \underbrace{\left(S_{i,T}, A_{i,T}, R_{i,T}\right)}_{D_{i,T}}$$

$D_{i,t} \triangleq (S_{i,t}, A_{i,t}, R_{i,t})$

Individual RL Algorithms

→ Algorithm Update
→ Data Collection



Dependence Within a User

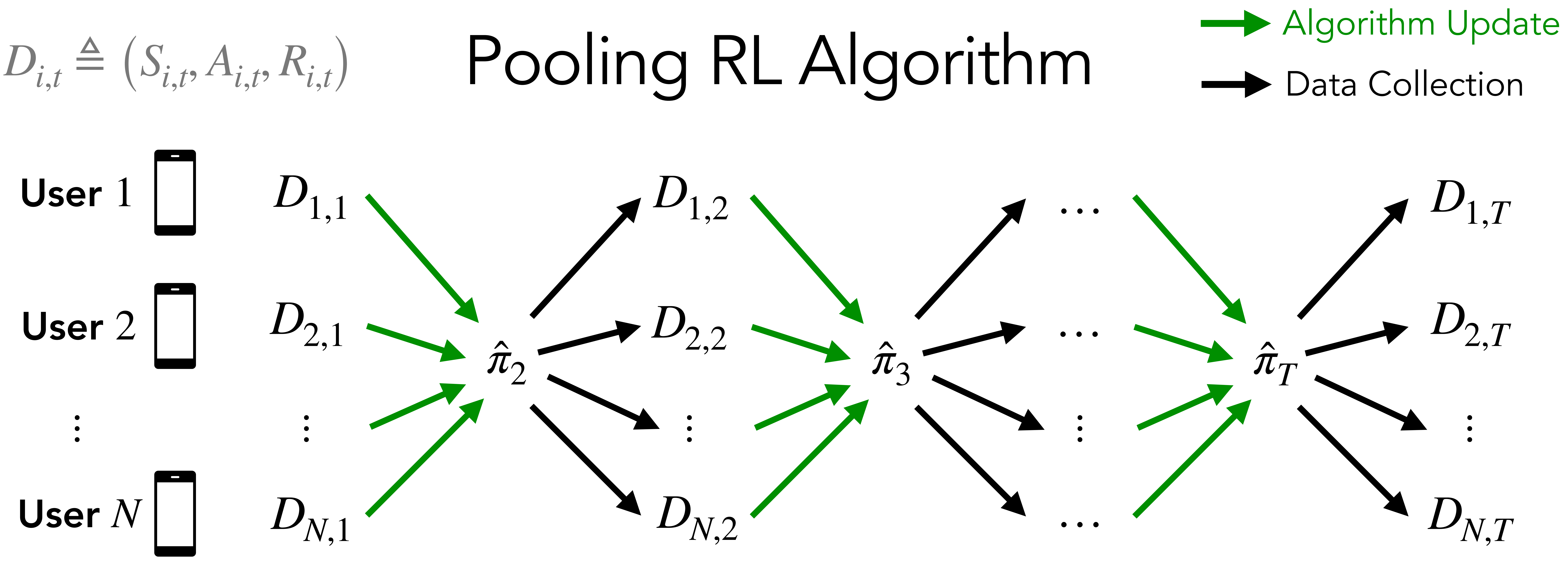
User states/rewards can be dependent over time

Limitations

Rewards are noisy and few decision times per user → slow learning

Pooling RL Algorithm

$D_{i,t} \triangleq (S_{i,t}, A_{i,t}, R_{i,t})$



Dependence Within a User

User states/rewards can be dependent over time

Dependence Between Users

Due to use of pooling algorithm

Inferential Goal

Treatment effect parameter $\theta_1^\star$

Parameters in an outcome model

- **Linear Model:**

$$\mathbb{E}\left[R_{i,t} \mid D_{i,1:t-1}, S_{i,t}, A_{i,t}\right] = \phi\left(D_{i,1:t-1}, S_{i,t}\right)^\top \theta_0^\star + A_{i,t} S_{i,t}^\top \theta_1^\star$$

- **Logistic Model:**

$$\mathbb{E}\left[R_{i,t} \mid D_{i,1:t-1}, S_{i,t}, A_{i,t}\right] = \left[1 + \exp\left\{\phi\left(D_{i,1:t-1}, S_{i,t}\right)^\top \theta_0^\star + A_{i,t} S_{i,t}^\top \theta_1^\star\right\}\right]^{-1}$$

General Case

$$\theta^\star \triangleq \operatorname{argmin}_\theta \mathbb{E}^\star\left[\ell_{\theta^\star}(D_{i,1:T})\right]$$

Typical Approach to Forming Estimators

Estimator $\hat{\theta}$ minimizes empirical loss:

$$\hat{\theta} \triangleq \operatorname{argmin} \frac{1}{T} \sum_{t=1}^{T} \ell_{\theta}(D_{i,1:T})$$

Examples

- Sample mean
- Least squares
- Logistic regression
- Maximum likelihood

Typical Approach to Forming Estimators

Estimator $\hat{\theta}$ minimizes empirical loss:

$$\hat{\theta} \triangleq \operatorname{argmin} \frac{1}{T} \sum_{t=1}^{T} \ell_\theta(D_{i,1:T})$$

Under certain assumptions on the adaptive policies

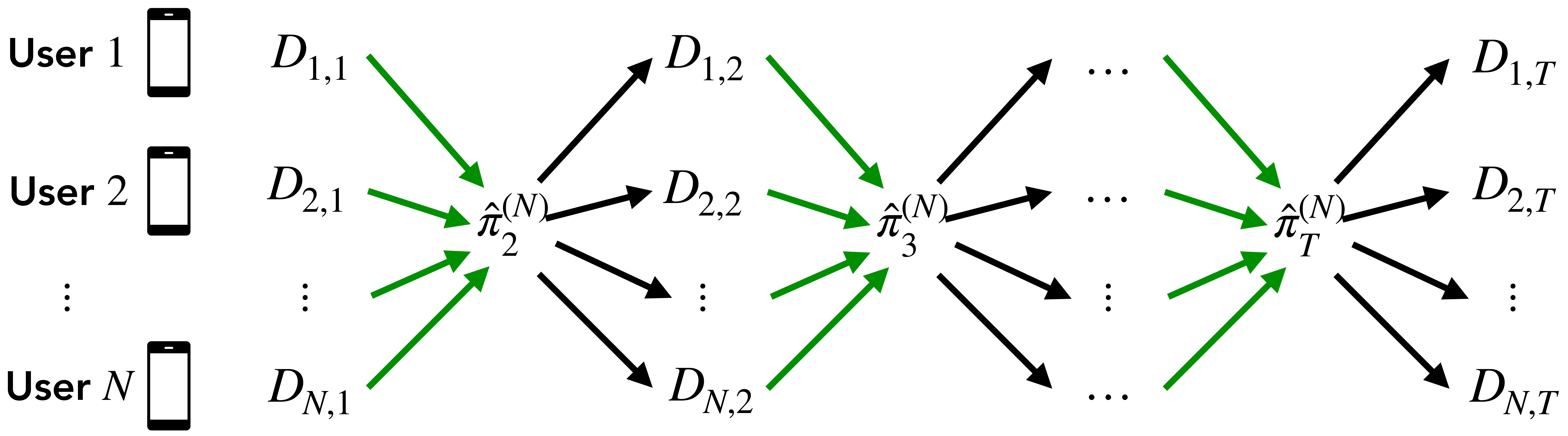
- Standard estimators are asymptotically normal
- However, common variance estimators inaccurate

$D_{i,t} \triangleq \left(S_{i,t}, A_{i,t}, R_{i,t}\right)$

Pooling RL Algorithm

Algorithm Update

Data Collection



For each $\hat{\pi}_t^{(N)}$ as $N \to \infty$,
$\hat{\pi}_t^{(N)} \to \pi_t^{\star}$ (limiting policy)

$$\hat{\pi}_t^{(N)}(s) = \mathbb{P}\left(A_{i,t} = 1 \,\middle|\, \left\{D_{i,1:t-1}\right\}_{i=1}^{N}, S_{i,t} = s\right)$$

Parametric Policy Classes

Policy Class: $\left\{\pi(\,\cdot\,;\beta)\right\}_{\beta\in\mathbb{R}^d}$

- Estimated policy: $\hat{\pi}_t^{(N)}(s) \triangleq \pi\left(s;\hat{\beta}_{t-1}^{(N)}\right)$
- Limiting policy: $\pi_t^{\star}(s) \triangleq \pi\left(s;\beta_{t-1}^{\star}\right)$

Form $\hat{\beta}_{t-1}^{(N)}$ with $\left\{D_{i,1:t-1}\right\}_{i=1}^{N}$

(e.g. estimate of reward model parameters)

Parametric Policy Classes

Policy Class: $\left\{\pi(\,\cdot\,;\beta)\right\}_{\beta\in\mathbb{R}^d}$

- Estimated policy: $\hat{\pi}_t^{(N)}(s) \triangleq \pi\left(s;\hat{\beta}_{t-1}^{(N)}\right)$
- Limiting policy: $\pi_t^{\star}(s) \triangleq \pi\left(s;\beta_{t-1}^{\star}\right)$

Form $\hat{\beta}_{t-1}^{(N)}$ with $\left\{D_{i,1:t-1}\right\}_{i=1}^{N}$

(e.g. estimate of reward model parameters)

Key Assumptions

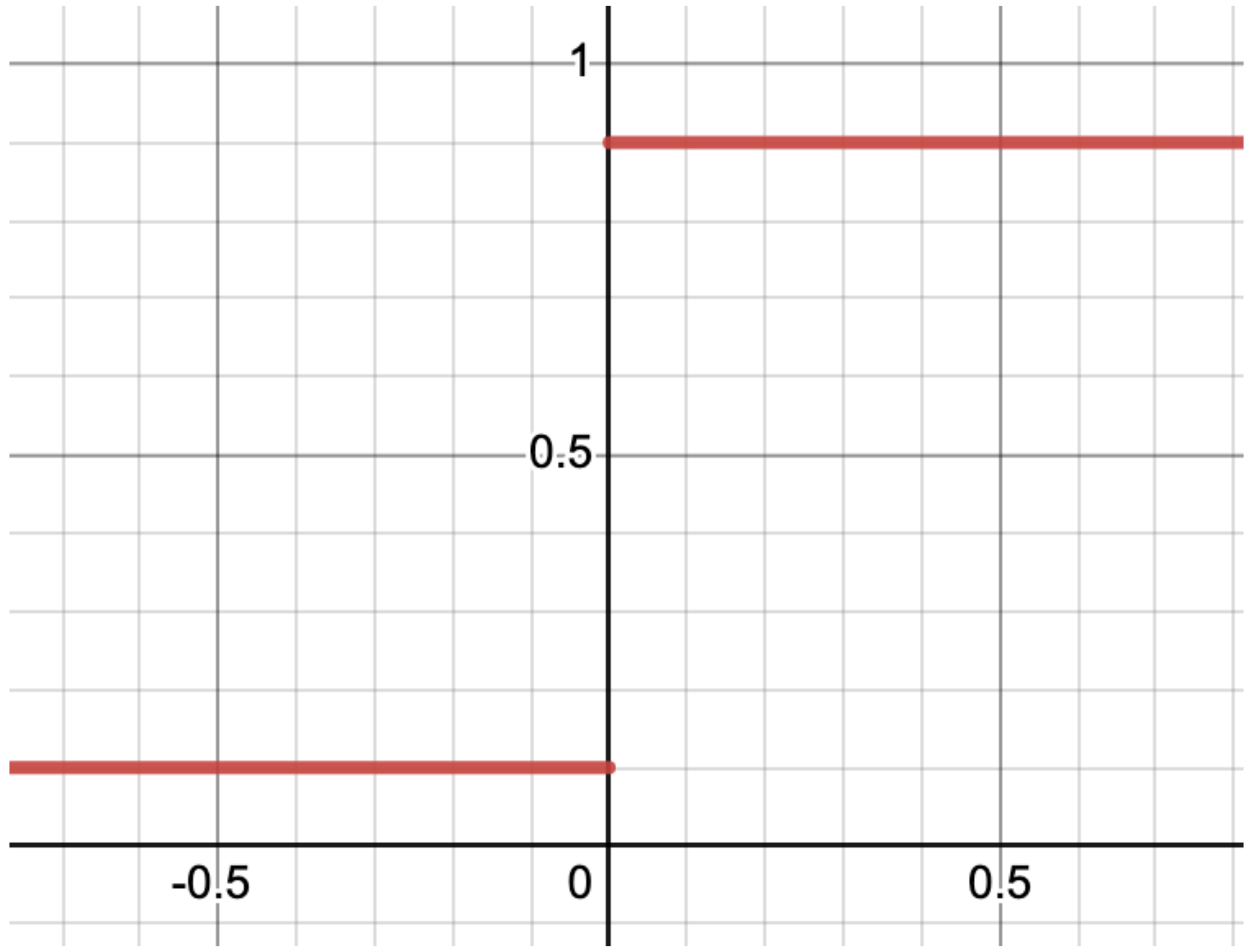
1. Convergence of $\hat{\beta}_t^{(N)} \xrightarrow{P} \beta_t^{\star}$ (for each t)
2. Policy class $\left\{\pi(\,\cdot\,;\beta)\right\}_{\beta\in\mathbb{R}^d}$ is smooth in β (Lipschitz)

What probability should the limiting policy send a message?

Maximize Rewards

$$\pi^\star(s) = \mathbf{1}\{\text{Treatment Effect}(s) > 0\}$$

Probability of Sending a Message

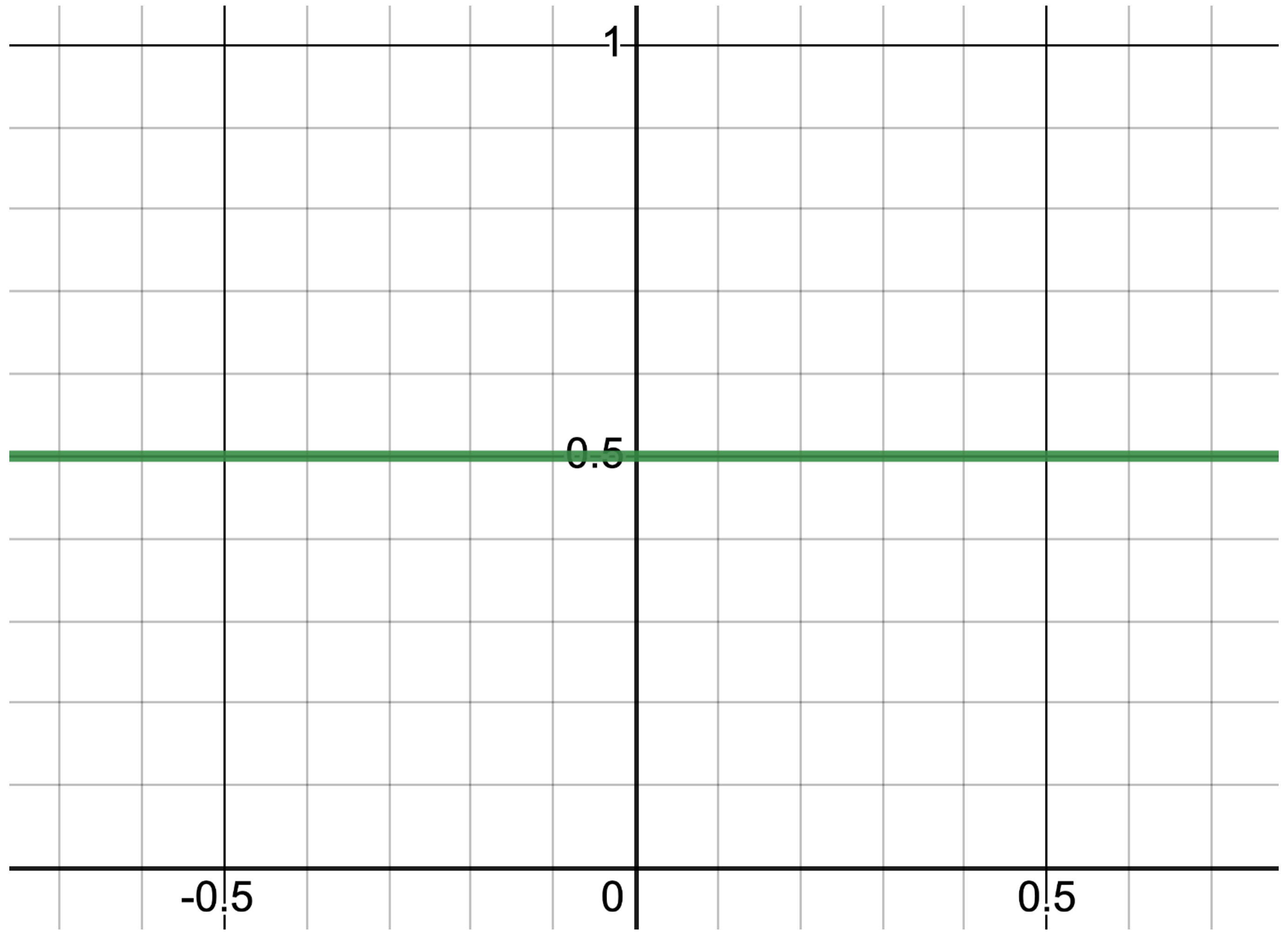


Treatment Effect in State s

Accurately Infer Treatment Effects

$$\pi^\star(s) = 0.5$$

Probability of Sending a Message



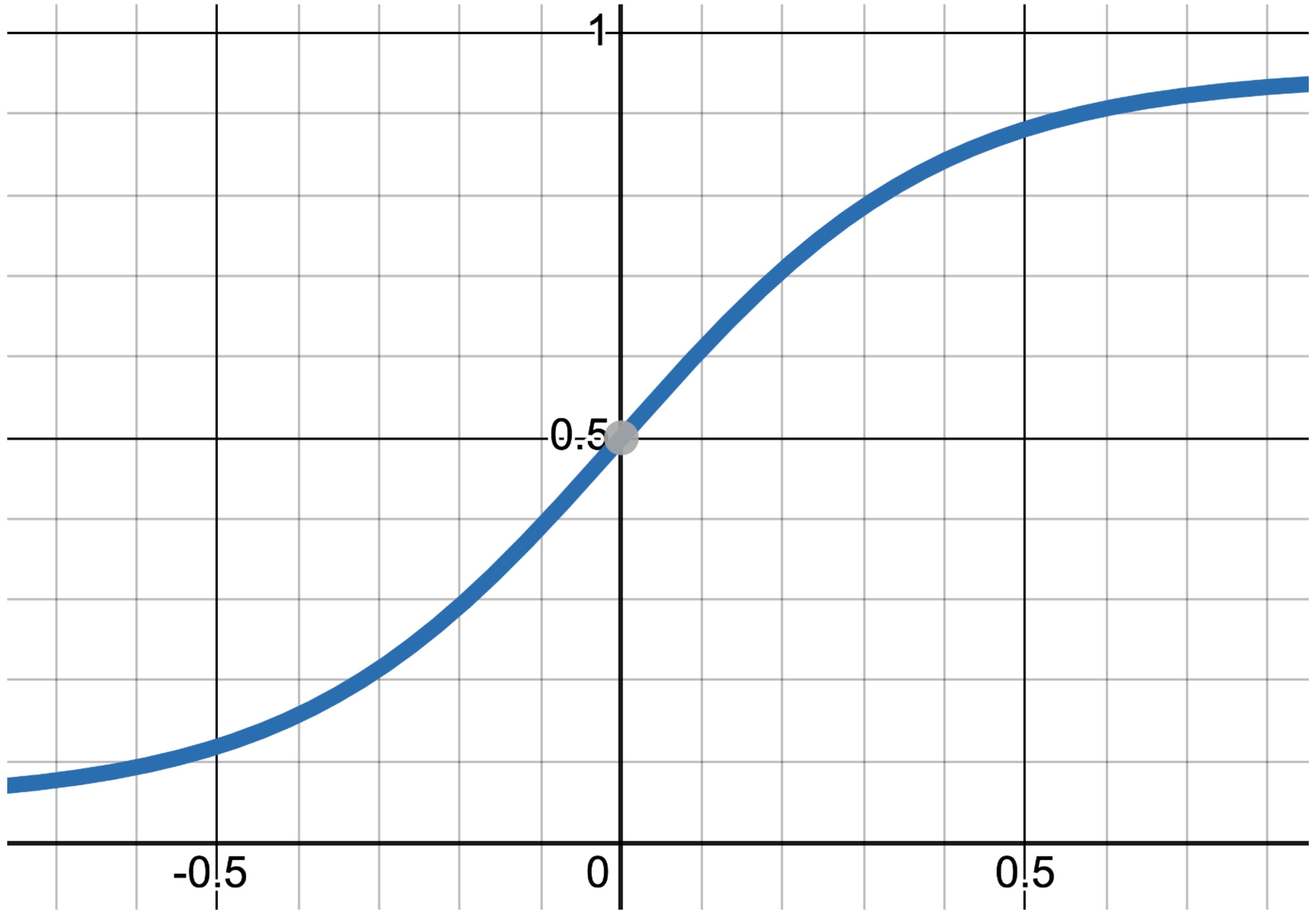
Treatment Effect in State s

What probability should the limiting policy send a message?

Balance Maximizing Rewards and Inferring Treatment Effects

$$\pi^{\star}(s) = \text{Softmax}\big(\text{Treatment Effect}(s)\big)$$

Probability of Sending a Message



Treatment Effect in State s

No longer have issue of unstable learned policies from taking a "hardmax"

Inference Challenges

(1) Dependencies both **within** and **between** users

(2) Error of $\hat{\theta}$ implicitly depends on how the algorithm forms and updates policies $\hat{\pi}_t$

Coverage of 95% Confidence Intervals for Treatment Effect

$\hat{\theta}$ Variance Estimators	$N = 50$	$N = 100$
Standard Sandwich	75.8%	77.6%
"Adaptive" Sandwich	95.4%	96.5%

Adaptive Sandwich Variance (Result Summary)

For **longitudinal data** collected by a particular class of **pooled RL algorithms**, under regularity conditions,

$$\sqrt{N}(\hat{\theta} - \theta^{\star}) \overset{D}{\not\to} \mathcal{N}(0, \underbrace{\Sigma}_{\text{Typical Variance (no RL)}})$$

Zhang, Janson, & Murphy, 2023
Under submission

Adaptive Sandwich Variance (Result Summary)

For **longitudinal data** collected by a particular class of **pooled RL algorithms**, under regularity conditions,

$$\sqrt{N}(\hat{\theta} - \theta^{\star}) \overset{D}{\nrightarrow} \mathcal{N}(0, \underbrace{\Sigma}_{})$$

Typical Variance (no RL)

$$\sqrt{N}(\hat{\theta} - \theta^{\star}) \overset{D}{\rightarrow} \mathcal{N}(0, \underbrace{\Sigma^{\text{adapt}}}_{})$$

Correction in Variance Due to Pooled RL Algorithm

Zhang, Janson, & Murphy, 2023
Under submission

Adaptive Sandwich Variance (Result Summary)

For **longitudinal data** collected by a particular class of **pooled RL algorithms**, under regularity conditions,

$$\sqrt{N}(\hat{\theta} - \theta^\star) \overset{D}{\nrightarrow} \mathcal{N}(0, \underbrace{\Sigma}_{})$$

Typical Variance (no RL)

$$\sqrt{N}(\hat{\theta} - \theta^\star) \overset{D}{\rightarrow} \mathcal{N}(0, \underbrace{\Sigma^{\text{adapt}}}_{})$$

Correction in Variance Due to Pooled RL Algorithm

Zhang, Janson, & Murphy, 2023

Under submission

Impact of Adaptive Sandwich Variance Approach

Enables the use of pooling RL algorithms in digital intervention studies

Oralytics:
Oral Health Coaching

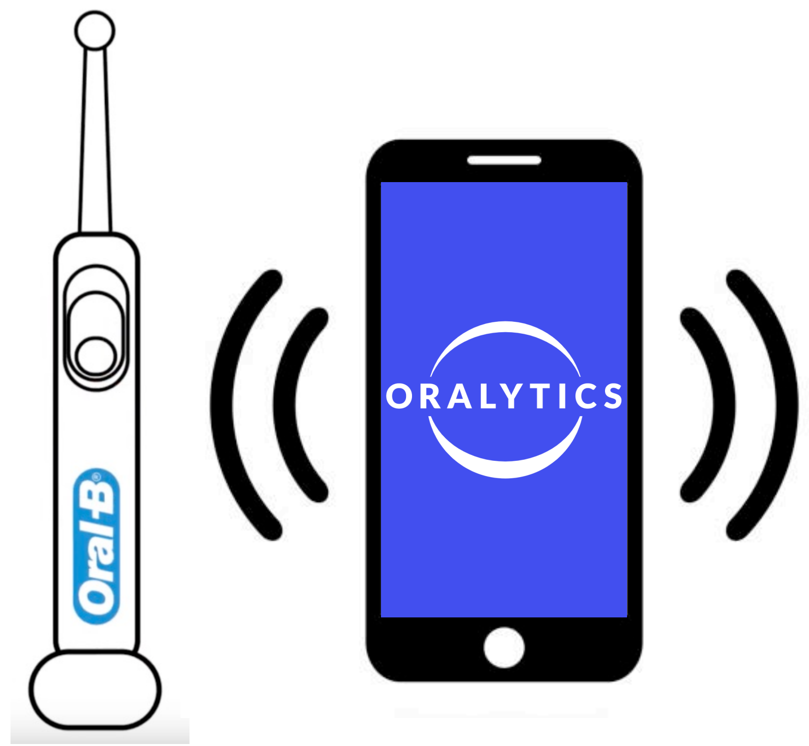
MiWaves:
Curbing Adolescent Marijuana Use

Oralytics: Designed RL Algorithm with Interdisciplinary Team

Pre-Implementation Guidelines for Online RL for Digital Interventions
Algorithms 2022 (Oral Presentation at RLDM 2022)
Trella, **Zhang**, Nahum-Shani, Shetty, Doshi-Velez. & Murphy

Reward Design for an Online RL Algorithm to Support Oral Self-Care
Innovative Applications of AI, 2023
Trella, **Zhang**, Nahum-Shani, Shetty, Doshi-Velez, & Murphy

Our RL algorithm is currently in the field!



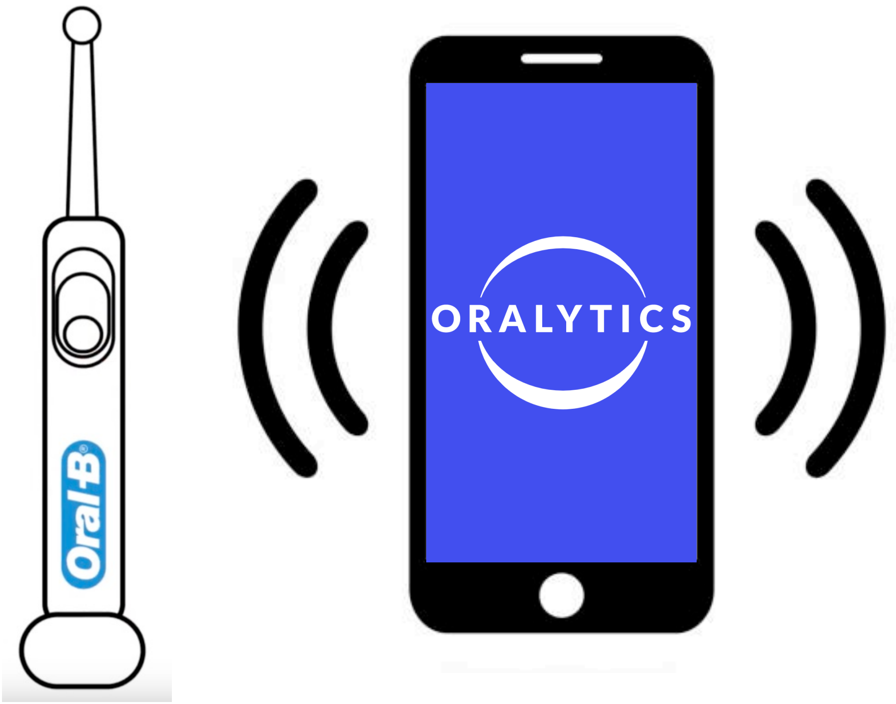
Conclusion

Summary

Part 1:
Contextual Bandit Setting

- Standard estimators **asymptotically non-normal** due to instability in adaptive policies
- **Adaptively weighted estimators** preserve asymptotic normality

Part 2:
Longitudinal Data Setting



- Using data from **"smooth"** adaptive policies, standard estimators are still asymptotically normal
- Need to **adjust variance estimator** to account for adaptive sampling

Future Work / Open Questions

Next Steps / Direct Extensions

- Software Package
- Incremental recruitment

Related Open Questions

- Different asymptotic regimes
- Randomization based inference
- Incorporating observational data and/or predictions from high dimensional ML models
- Other forms of pooling: limited resource allocation, partial pooling

Acknowledgements

Advisors

Lucas Janson

Susan Murphy

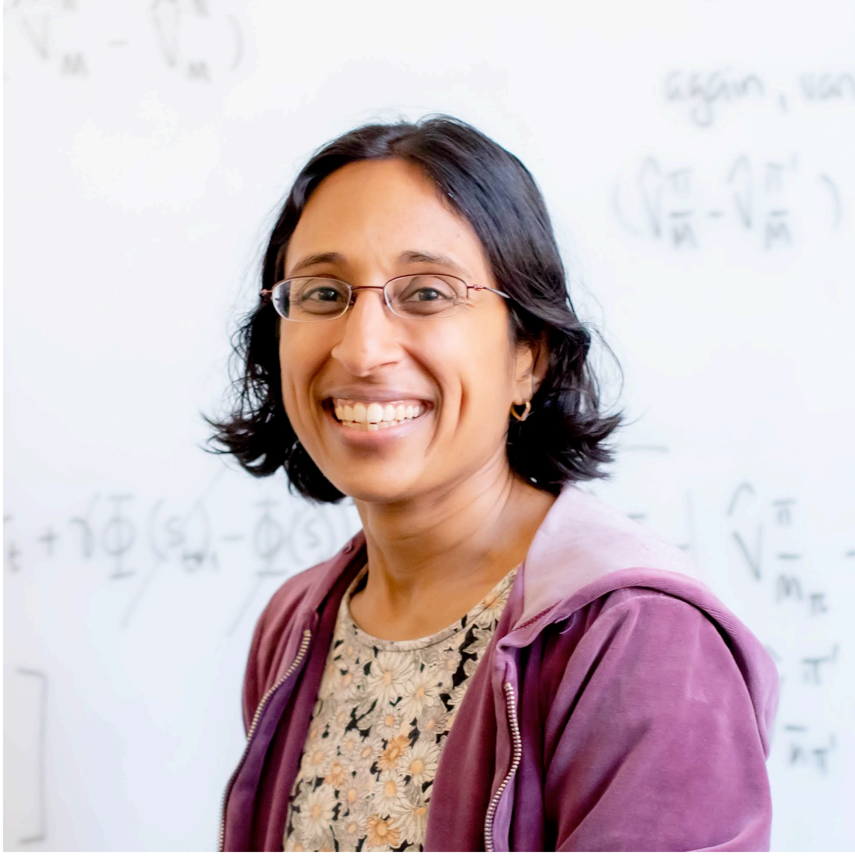
Collaborators!

Anna Trella

Raaz Dwivedi

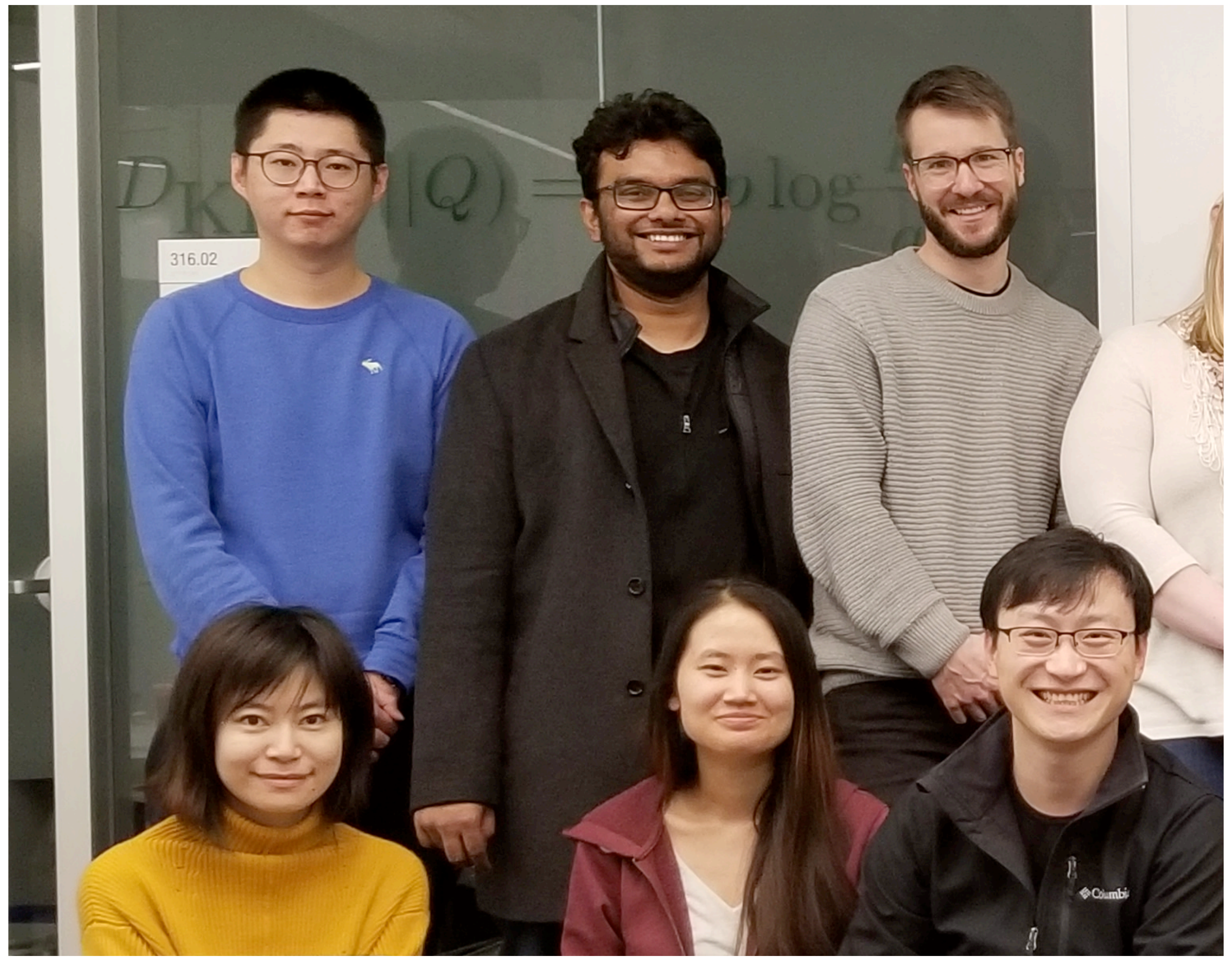
Inbal Nahum-Shani

Vivek Shetty

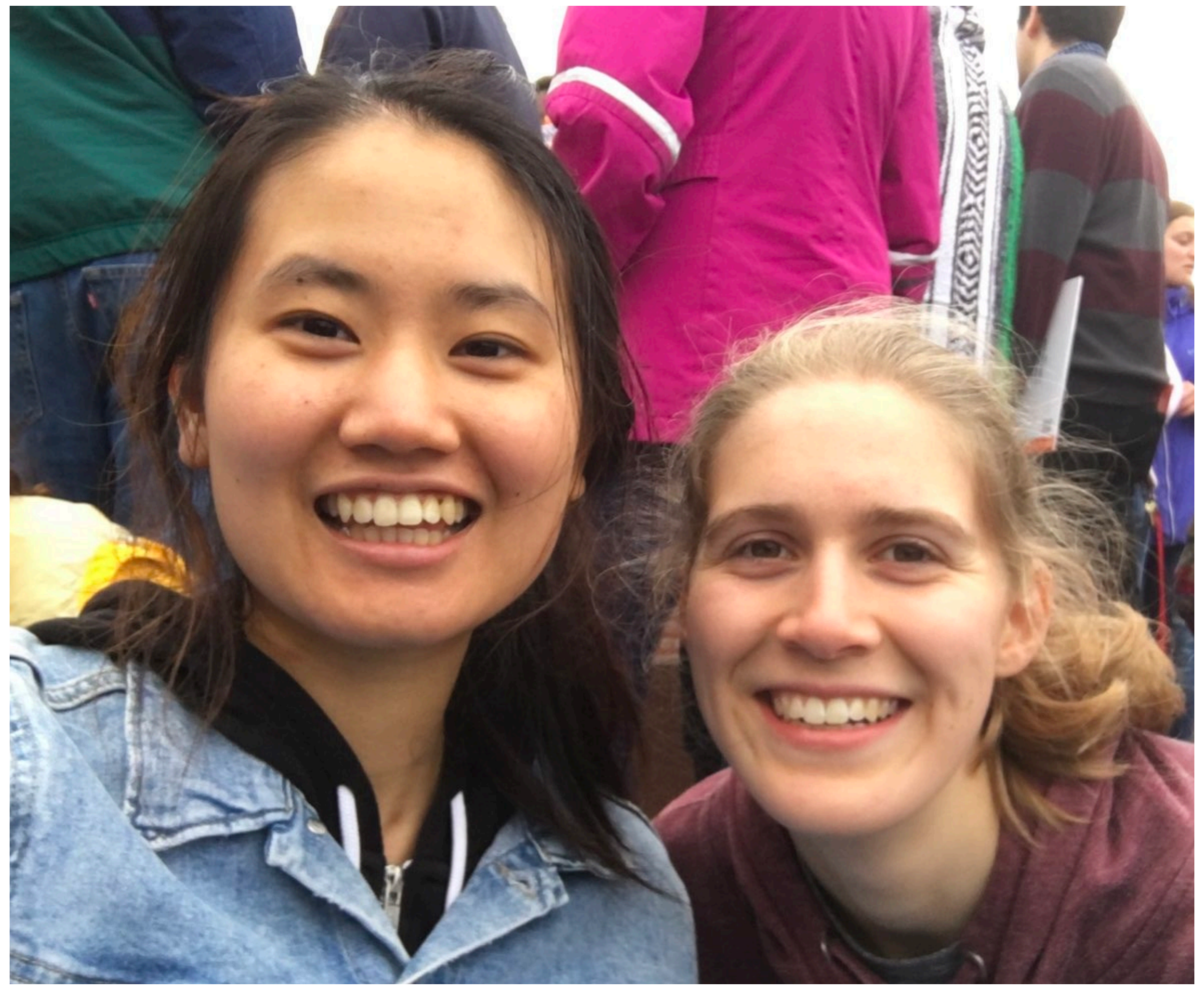


Finale Doshi-Velez

Stat RL Lab Friends



Friends



Family

Backup Slides

Oralytics: The State of Dental Health

Oral diseases are largely preventable through regular brushing and flossing

- 5-10% of healthcare budgets in industrialized countries are spent on treating dental cavities

World Health Organization

- Nearly one-fifth of U.S. adults 65 or older have lost all their teeth

CDC
Centers for Disease Control and Prevention

Adaptive Sandwich Variance

$$\sqrt{n}\left(\hat{\theta}^{(n)} - \theta^{\star}\right) \xrightarrow{D} \mathcal{N}\left(0,\ \ddot{L}^{-1}\Sigma^{\text{adapt}}\ \ddot{L}^{-1}\right)$$

$$\Sigma^{\text{adapt}} = \mathbb{E}_{\pi^{\star}}\left[\left\{\dot{\ell}\left(D_{i,1:T};\theta^{\star}\right) + \underbrace{\dot{L}^{-1}\sum_{t=1}^{T-1} f_t\left(D_{i,1:t};\beta_t^{\star}\right)}_{\text{Correction in Variance Due to Pooled RL Algorithm}}\right\}^{\otimes 2}\right]$$

f_t **given in paper:** Statistical Inference After Adaptive Sampling for Longitudinal Data (https://arxiv.org/abs/2202.07098)