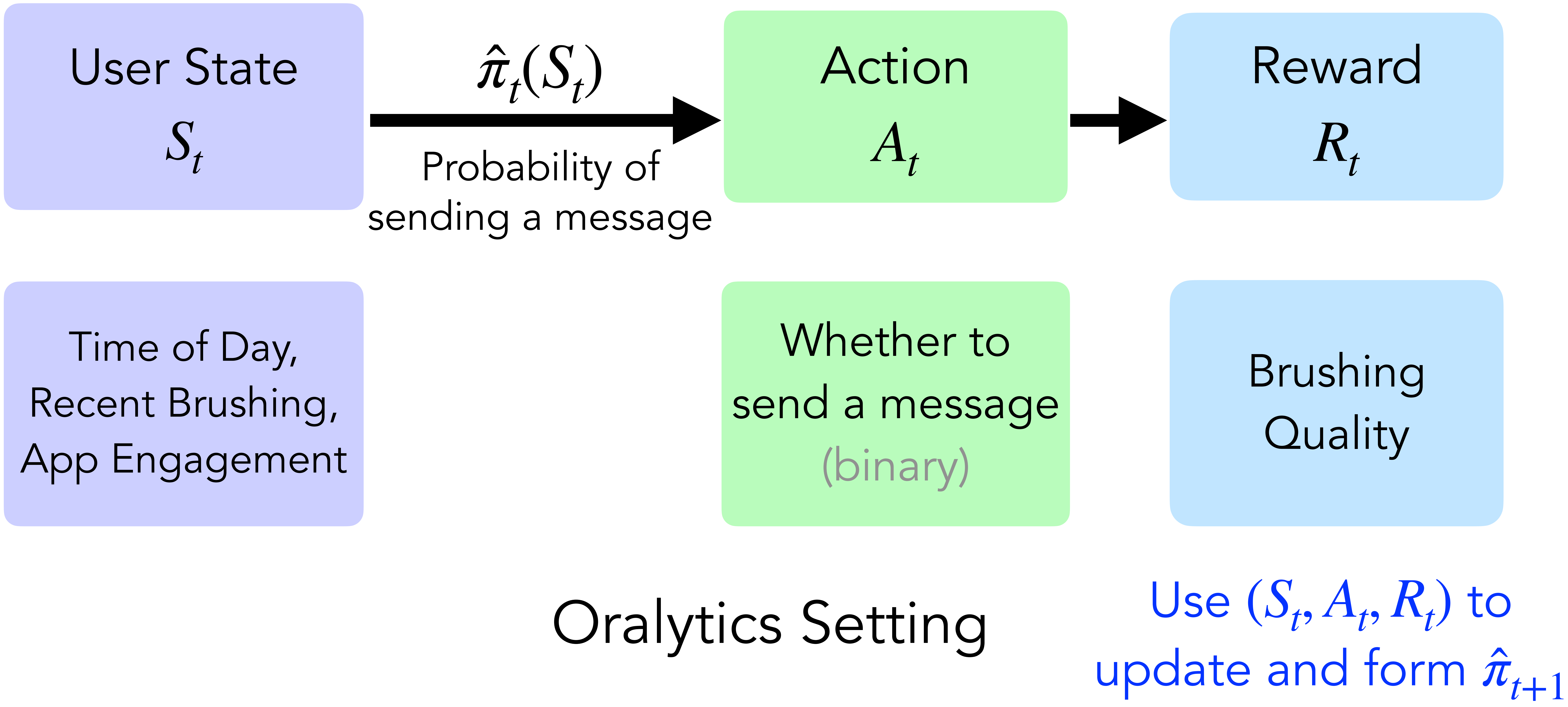
Software Package for Inference after Pooling

How to Interface with the RL Algorithm

Kelly Zhang and Nowell Closser

Motivation and Challenges Due to Using Pooling RL Algorithms

Online Reinforcement Learning (RL)



MRT Study Data

- **Total Decision Times:** T
- **Number of users:** N
- **Data Collected After Study:** For each user $i \in [1: N]$,

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

- **User History:** $H_{i,t} = \left\{D_{i,1}, D_{i,2}, D_{i,3}, \ldots, D_{i,t}\right\}$

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

Individual RL Algorithms

$\longrightarrow$ Algorithm Update

$\longrightarrow$ Data Collection

$$
\begin{array}{lcccccccccccc}
\textbf{User } 1 & D_{1,1} & \longrightarrow & \hat{\pi}_{1,2} & \longrightarrow & D_{1,2} & \longrightarrow & \hat{\pi}_{1,3} & \longrightarrow & \dots & \longrightarrow & \hat{\pi}_{1,T} & \longrightarrow & D_{1,T} \\
\textbf{User } 2 & D_{2,1} & \longrightarrow & \hat{\pi}_{2,2} & \longrightarrow & D_{2,2} & \longrightarrow & \hat{\pi}_{2,3} & \longrightarrow & \dots & \longrightarrow & \hat{\pi}_{2,T} & \longrightarrow & D_{2,T} \\
\vdots & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\
\textbf{User } N & D_{N,1} & \longrightarrow & \hat{\pi}_{N,2} & \longrightarrow & D_{N,2} & \longrightarrow & \hat{\pi}_{N,3} & \longrightarrow & \dots & \longrightarrow & \hat{\pi}_{N,T} & \longrightarrow & D_{N,T}
\end{array}
$$

Dependence Within a User

User states/rewards can be dependent over time

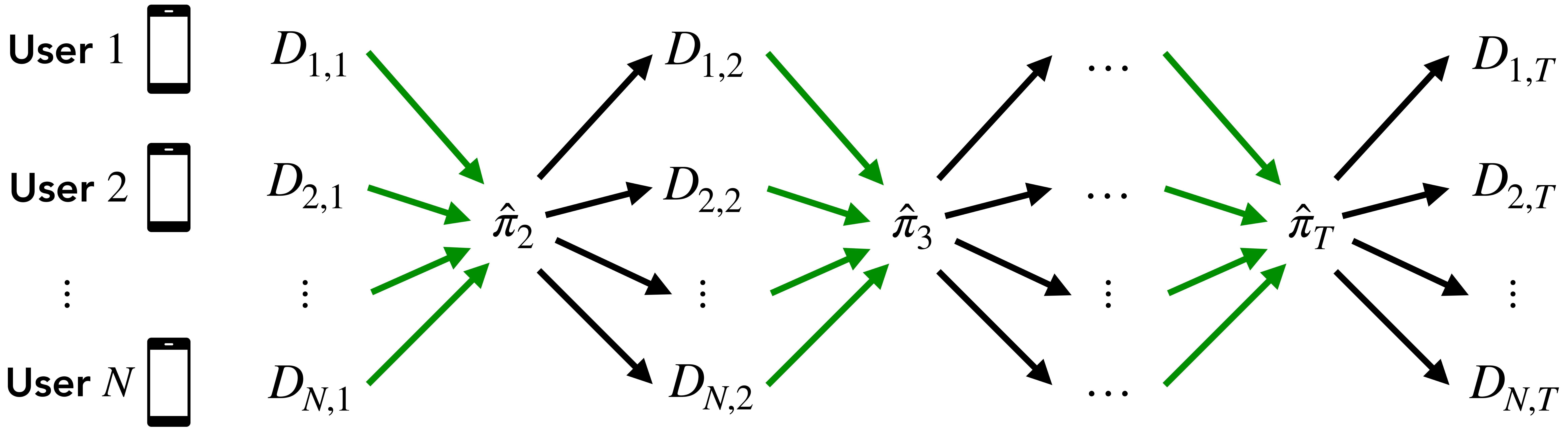
Limitations

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

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

Pooling RL Algorithm

Algorithm Update
Data Collection



Dependence Within a User

User states/rewards can be dependent over time

Dependence Between Users

Due to use of pooling algorithm

Inference After Pooling Approach Overview

Pooling Algorithm

- At each decision time t, use **all past user data** to form a statistic $\hat{\beta}_{t-1}^{(N)}$ (e.g. estimator for parameters in a reward model)
- $\hat{\pi}_t(s) = \pi_t\left(s; \hat{\beta}_{t-1}^{(N)}\right)$ is pooled policy at time t
- Select action $A_{i,t} \Big| S_{i,t}, H_{1:n,t-1} \sim \text{Bernoulli}\left(\hat{\pi}_t(S_{i,t})\right)$

What is $\hat{\beta}_t$ in general?

$$\phi_t(H_t; \beta) = (R_t - \beta_a^\top f(S_t) - \beta_b^\top A_t g(S_t)) \begin{bmatrix} f(S_t) \\ A_t g(S_t) \end{bmatrix}$$

$\hat{\beta}_t \in \mathbb{R}^{p_\beta}$ is the solution to an estimating equation for

$$0 = \frac{1}{N} \sum_{i=1}^{N} \phi_t \left(H_{i,t}; \hat{\beta}_{1:t} \right) \in \mathbb{R}^{p_\beta}$$

e.g. minimizer of a loss function, least squares, MLEs, etc.

Theory applies when: $\hat{\beta}_t^{(N)} \to \beta_t^\star$ as $N \to \infty$

where $0 = \mathbb{E}_\star \left[\phi_t \left(H_{i,t}; \beta_{1:t}^\star \right) \right] \in \mathbb{R}^{p_\beta}$

Inference Approach

- **Inferential Goal:**

$$0 = \mathbb{E}_\star \left[\psi \left(H_{i,T}; \theta^\star, \beta^\star_{1:T-1} \right) \right] \in \mathbb{R}^{p_\theta}$$

- **Estimator:**

$$0 = \frac{1}{N} \sum_{i=1}^{N} \psi \left(H_{i,T}; \hat{\theta}, \hat{\beta}_{1:T-1} \right) \in \mathbb{R}^{p_\theta}$$

e.g. minimizer of a loss function, least squares, MLEs, etc.

Asymptotic Normality Result (T = 3 Case)

$$\sqrt{n}\begin{pmatrix}\hat{\beta}_1 - \beta_1^\star \\ \hat{\beta}_2 - \beta_2^\star \\ \hat{\theta} - \theta^\star\end{pmatrix} \xrightarrow{D} \mathcal{N}\left(0, B^{-1}\Sigma(B^{-1})^\top\right)$$

$$\Sigma = \mathbb{E}_\star\left[\begin{pmatrix}\phi_1(H_{i,1};\beta_1^\star) \\ \phi_2(H_{i,2};\beta_{1:2}^\star) \\ \psi(H_{i,3};\theta^\star,\beta_{1:2}^\star)\end{pmatrix}^{\otimes 2}\right] = \mathbb{E}_\star\left[\begin{pmatrix}\phi_{i,1}(\beta_1^\star) \\ \phi_{i,2}(\beta_{1:2}^\star) \\ \psi_i(\theta^\star,\beta_{1:2}^\star)\end{pmatrix}^{\otimes 2}\right]$$

Asymptotic Normality Result (T = 3 Case)

$$\sqrt{n}\begin{pmatrix}\hat{\beta}_1 - \beta_1^\star \\ \hat{\beta}_2 - \beta_2^\star \\ \hat{\theta} - \theta^\star\end{pmatrix} \xrightarrow{D} \mathcal{N}\left(B^{-1}\Sigma(B^{-1})^\top\right)$$

$$B = \frac{\partial}{\partial(\beta_1, \beta_2, \theta)}\mathbb{E}_\star\left[\begin{matrix}\phi_{i,1}(\beta_1) \\ W_{i,2}^\star(\beta_1)\phi_{i,2}(\beta_{1:2}) \\ W_{i,2}^\star(\beta_1)W_{i,3}^\star(\beta_2)\psi_i(\theta, \beta_{1:2})\end{matrix}\right]\Bigg|_{\beta_1^\star, \beta_2^\star, \theta^\star}$$

$$W_{i,t}^\star(\beta_{t-1}) = \left(\frac{\pi_t(S_{i,t}; \beta_{t-1})}{\pi_t(S_t; \beta_{t-1}^\star)}\right)^{A_{i,t}}\left(\frac{1 - \pi_t(S_{i,t}; \beta_{t-1})}{1 - \pi_t(S_t; \beta_{t-1}^\star)}\right)^{1-A_{i,t}}$$

Rewriting the "bread" part B

$$\mathbb{E}_\star \begin{bmatrix} \frac{\partial}{\partial \beta_1}\phi_1(\beta_1) & 0 & 0 \\ \left[\frac{\partial}{\partial \beta_1} W_2^\star(\beta_1)\right]\phi_2 + \frac{\partial}{\partial \beta_1}\phi_2(\beta_{1:2}) & \frac{\partial}{\partial \beta_2}\phi_2(\beta_{1:2}) & 0 \\ \left[\frac{\partial}{\partial \beta_1} W_2^\star(\beta_1)\right]\psi + \frac{\partial}{\partial \beta_1}\psi(\beta_{1:2},\theta) & \left[\frac{\partial}{\partial \beta_2} W_3^\star(\beta_2)\right]\psi + \frac{\partial}{\partial \beta_2}\psi(\beta_{1:2},\theta) & \frac{\partial}{\partial \theta}\psi(\beta_{1:2},\theta) \end{bmatrix} \Bigg|_{\beta_{1:2}^\star, \theta^\star}$$

$$W_t^\star(\beta_{t-1}) = \left(\frac{\pi_t(S_t;\beta_{t-1})}{\pi_t(S_t;\beta_{t-1}^\star)}\right)^{A_t} \left(\frac{1-\pi_t(S_t;\beta_{t-1})}{1-\pi_t(S_t;\beta_{t-1}^\star)}\right)^{1-A_t}$$

Software and Computational Challenges

Estimating the Matrix B

$$\frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix} \frac{\partial}{\partial\beta_1}\phi_{i,1} & 0 & 0 \\ \left[\frac{\partial}{\partial\beta_1}W_{i,2}(\beta_1)\right]\phi_{i,2}+\frac{\partial}{\partial\beta_1}\phi_{i,2} & \frac{\partial}{\partial\beta_2}\phi_{i,2} & 0 \\ \left[\frac{\partial}{\partial\beta_1}W_{i,2}(\beta_1)\right]\psi_i+\frac{\partial}{\partial\beta_1}\psi_i & \left[\frac{\partial}{\partial\beta_2}W_{i,3}(\beta_2)\right]\psi_i+\frac{\partial}{\partial\beta_2}\psi_i & \frac{\partial}{\partial\theta}\psi_i \end{bmatrix}\Bigg|_{\hat\beta_{1:2},\hat\theta}$$

$$W_{i,t}(\beta_{t-1}) = \left(\frac{\pi_t(S_{i,t};\beta_{t-1})}{\pi_t(S_{i,t};\hat\beta_{t-1})}\right)^{A_{i,t}}\left(\frac{1-\pi_t(S_{i,t};\beta_{t-1})}{1-\pi_t(S_{i,t};\hat\beta_{t-1})}\right)^{1-A_{i,t}}$$

$$\frac{\partial}{\partial\beta_{t-1}}\pi_t(S_{i,t};\beta_{t-1})$$

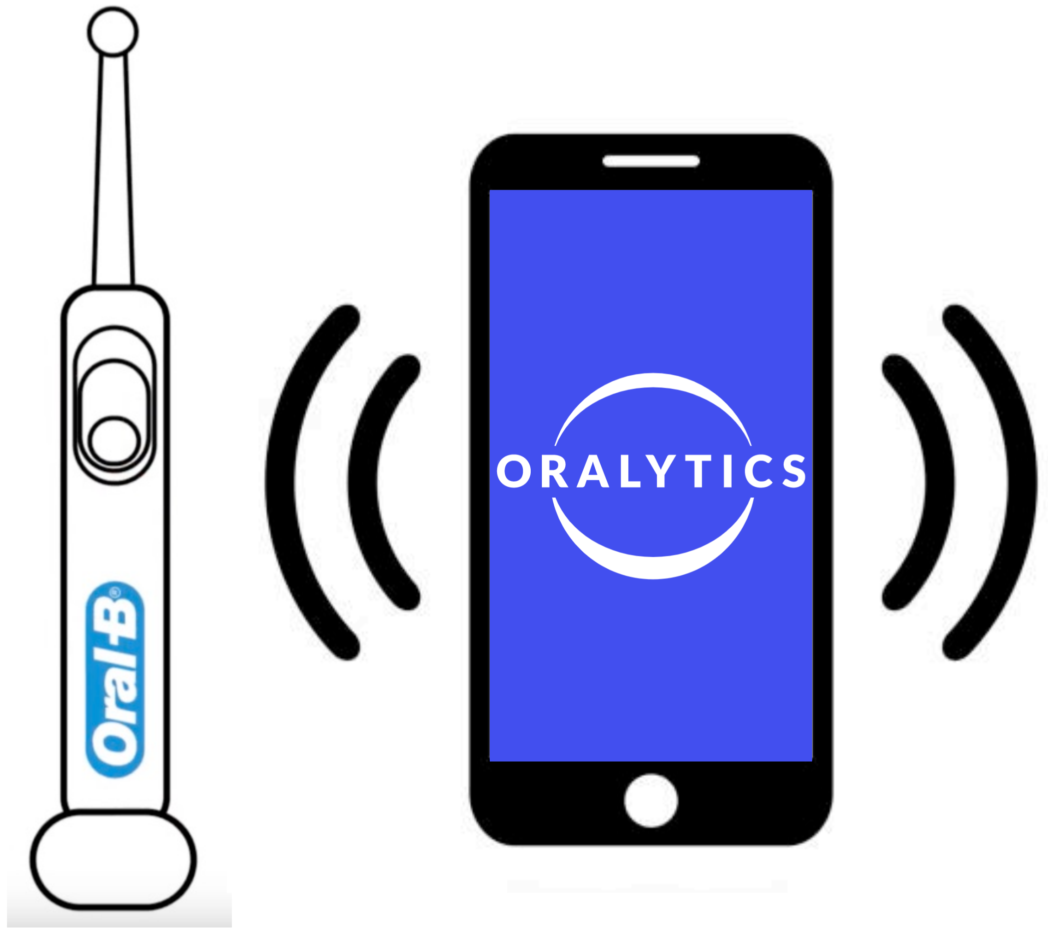
Considerations

1. **Quality of Variance Estimator**
    - Accuracy and Numerical Stability
2. **Computational Considerations**
    - How long to compute the standard errors from a run? (simulate thousands of runs)
3. **RL Algorithm Designer Experience**
    - Want to allow for flexibility in algorithm design
    - Want to make it easy and automatic for RL algorithm to provide necessary statistics
4. **Data Analyzer Experience**
    - Want to allow for a variety of analyses
    - Want to allow data analyzer to have little to no knowledge of RL algorithm to use

Approaches (go to following google doc):

https://docs.google.com/document/d/1NVAiaqv5fNhPUtkUMv1pPY55jO4jUDx0hdlKrJPy84Q/edit?usp=sharing

Backup Slides



UNIVERSITY OF MICHIGAN

UCLA School of Dentistry

Digital Oral Health Coaching

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

Online Reinforcement Learning (RL)

Maximize:
User Benefit

Digital Intervention Study Design Objectives

Within-Study Personalization

Maximize User Benefit

- Send messages at opportune moments

Use Online RL Algorithms

$$\mathbb{E}\left[\sum_{t=1}^{T} R_t\right]$$

After-Study Analyses

Digital Intervention Study Design Objectives

Within-Study Personalization

Maximize User Benefit

- Send messages at opportune moments

Use Online RL Algorithms

$$\mathbb{E}\left[\sum_{t=1}^{T} R_t\right]$$

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

$$\mathbb{E}\left[\sum_{t=1}^{T} R_t\right]$$

After-Study Analyses

Confidence Intervals Critical for

- Replicable science
- Publishing and sharing results

Infer Treatment Effects

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