Robust tests in online decision-making: testing the utility of data collected by wearables

Jane Paik Kim
Stanford University
janepkim@stanford.edu

February 25, 2019
Overview

1. Background: Mobile health in mental health care
2. Apple study at Stanford and practical issues
3. Online decision making algorithms
4. Assessing the utility of context variables in the actor-critic bandit
5. Robust tests for testing the utility of variables under model misspecification
Mobile health in Mental Health Today

Shift in mental health care delivery

- Apps need to adapt and engage
- Details of implementation (content, sequence, timing) matter
- Behavioral theory is limited in guiding the design of the decision rules

Figure: Nature, 2016
The digital health industry occupies a large part of the health care space and has been rapidly expanding

- mHealth apps for treatment of mental health offer various models of care
- NIMH’s categorization of functionality: self-management, cognition improvement, skills-training, social support, symptom tracking, and passive data collection
- Apps span all stages of care provision, from diagnosis, treatment, to maintenance
Projects in mental health

Self-monitoring and Cognitive Behavioral Therapy

- Eating disorders (Tregarthen et al. 2015)

- Bipolar disorders (Gliddon et al. 2018)

MoodSwings
An online self help program for Bipolar Disorder
Stanford Center for Digital Health: Apple Watch Seed Research Grant Program

Awarded 5 teams at Stanford, including us

Our Study: Improving Adherence Behaviors using Apple Watches

App goal: to promote adherence to a daily fitness goal (step count) among individuals in good health, by using push notifications delivered by the watch

Study aims

- Scope: Feasibility (n=300 watches)
- Generating a Dataset for Offline Policy Evaluation
Apple Watch Study

- At onboarding, users complete a baseline survey on the iPhone
  - Daily goal (i.e. step count)
  - Demographics
  - Selection of values from a pre-populated list
  - User selects an image best representing their motivating value
- Continuous data collection via the Apple Watch
- Follow-up for 8 weeks
Apple Watch Study Design

Up to 4 times a day, the app delivers these message types

2. Values: “The things that we value the most are the reasons we often make the choices we do.”
3. Image selected:

4. No message

Currently completing app development phase; starting recruitment phase
We use the HealthKit API to obtain health-related data collected by the phone and watch. The data types are:

- **Characteristic data**
- **Sample data**: information about the user’s health at any moment of time
- **Workout data**: data about the fitness, workout activity
- **Caveat**: user has to grant permission to health related apps
Issues in the Apple Watch study that are shared in mobile health

- The frequency of data queries to the Health Kit API impacts the app performance.
- We send requests to the HealthKit API and we can customize the frequency at which we send the requests.
- If the queries are too frequent, the app processing time slows down the app from the user perspective.
- The frequency of the data query and the reward feedback go hand in hand.
3 Data issues that need appropriate methods

- 1. Data query frequency. We observe reward feedback once a day, i.e. the daily step count. With 4 possible message choices, there are $4^4 = 256$ possible message sequences given in a day, so this causes a high dimensional action space.

- 2. Images are collected at baseline. This implies that the image can be used as factors- feature reduction techniques or contextual bandits beyond linear models (neural networks) may be needed.

- 3. In the Apple watch study, the question is whether the watch could be useful.

I'll focus on the last topic.
3. Assessing values of the context

In the Apple Watch study, one of the focal issues is whether the watch provides **useful** covariates.

- Health kit provides activity and health data
- In view of various costs, we should know if we are gathering information that is even useful
- Testing the utility of the variables obtained from the watch (self report or health-kit) plays an important role
Sequential decision making

The Apple Watch study is about delivering messages to motivate users to be active. Sequential decision making is required in an ‘online’ fashion.

- User-specific watch data is received. Based on these inputs, a sequence of messages is sent based on a policy.
- After messages are delivered, the users’ daily physical activity is observed.

The decision rule is revised to maximize the success of achieving their goal.
- Choices of messages are updated when the inputs accumulate.

**Figure 2.1**: Online decision algorithm.

At each time $t$, a greedy algorithm would generate an estimate $\hat{\theta}_k$ of the mean reward for each $k$th action, and select the action that attains the maximum among these estimates.

Suppose there are three actions with mean rewards $\hat{\theta}_3$. In particular, each time an action $k$ is selected, a reward of 1 is generated with probability $\hat{\theta}_k$. Otherwise, a reward of 0 is generated. The mean rewards are not known to the agent. Instead, the agent’s beliefs in any given time period about these mean rewards can be expressed in terms of posterior distributions. Suppose that, conditioned on the observed history $H_t \neq 1$, posterior distributions are represented by the probability density functions plotted in Figure 2.2. These distributions represent beliefs after the agent tries actions 1 and 2 one thousand times each, action 3 three times, receives cumulative rewards of 600, 400, and 1, respectively, and synthesizes these observations with uniform prior distributions over mean rewards of each action. They indicate that the agent is confident that mean rewards for actions 1 and 2 are close to their expectations of approximately 0.6 and 0.4. On the other hand, the agent is highly uncertain about the mean reward of action 3, though he expects 0.4.

The greedy algorithm would select action 1, since that offers the maximal expected mean reward. Since the uncertainty around this expected mean reward is small, observations are unlikely to change the expectation substantially, and therefore, action 1 is likely to be selected.
Multi-armed bandits

- Arms: $i = 1, 2, \cdots, N$.
- At time step $t = 1, 2, \cdots, T$, pull or choose arm $a(t) \in \{1, \cdots, N\}$.
- Observe reward at time $t$, $R_{a(t)}(t)$ where $R_i(t) \in \mathbb{R}$,
  \[ E(R_{a(t)}(t)) = m_{a(t)}. \]
  \[ E(R_i(t)) = m_i, \text{ and } m_{i^*} \geq m_j \text{ for all } j \neq i^*. \]
- Bandit feedback:
  \[ (R_1(t), R_2(t), \cdots R_{a(t)}-1(t), R_{a(t)}+1(t), \cdots, R_N(t)) \]
  are not observed.
- After collecting information, choose arm $a(t + 1) \in \{1, \cdots, N\}$.
- Goal: Minimize regret through repeated decision making,
  \[ \text{Regret}(T) \equiv E\left[ \sum_{t=1}^{T} m_{i^*} - m_{a(t)} \right] \]
Exploitation: select arms that appear best i.e. yields maximum expected reward according to the estimates

Exploration: experimentation to learn about their expected rewards

To maximize reward, a trade off is needed: both exploitation and exploration should be incorporated.

Unguided exploration strategies may lead to a cumulative regret linear to the number of trials ($T$)

Directed exploration methods (UCB and Thompson sampling) addresses the trade-off
Contextual Bandits (Woodroofe, 1979; Sarkar, 1991)

- Context of arm $i$ at time $t$: $s_i(t) \in \mathbb{R}^d$.
- Each arm yields a user-specific reward:

$$R_i(t) := s_i(t)^T \mu + \epsilon_i(t)$$

$\mu \in \mathbb{R}^d$ are unknown.
- Given $H_{t-1}$, $\epsilon_i(t)$ is $\sigma$-sub-Gaussian, i.e., for $\forall \alpha \in \mathbb{R}$,

$$\mathbb{E}\left[\exp(\alpha \epsilon_i(t))\bigg| H_{t-1}\right] \leq \exp\left(\frac{\alpha^2 \sigma^2}{2}\right).$$

where $H_{t-1} = \left\{a(\tau), r_a(\tau), \{s_i(\tau)\}_{i=1}^N, \tau = 1, \cdots, t-1, \{s_i(t)\}_{i=1}^N\right\}$,

$$\mathbb{E}\left[R_i(t)\big| \{s_i(t)\}_{i=1}^N, H_{t-1}\right] = s_i(t)^T \mu,$$

- Without loss of generality, we assume $\|s_i(t)\| \leq 1$, $\|\mu\| \leq 1$. 

Jane Paik Kim (Stanford)
Value based approach: The goal is to learn $\mu$ in order to find the maximum of the expected reward

- Agent receives $(s_1(t), \cdots, s_N(t))$ as input.
- Let $a^*(t) = \arg\max_{1 \leq i \leq N} \{s_i^T(t)\mu\}$
- $\text{Regret}(T) \equiv E \left[ \sum_{t=1}^{n} s_{a^*(t)}^T \mu - s_{a(t)}^T \mu \right]$ The goal is to minimize the cumulative regret.
Alternatives to Value function approaches: Policy gradient approaches

- Note that the goal is to find a maximum of the conditional mean over the actions. The problem of finding the max becomes harder with a larger action space.
- This motivates using alternative approaches and directly finding the policy, or allocation probability.
- Policy gradient approaches explicitly learn the policy $\pi_\theta$ using a parameterized family of policies.
- Given $\pi_\theta$, find $\theta$ that maximizes the average reward.

$$V^*(\theta) = \int_{s \in S} d(s) \sum_{a \in A} E(R|s, a)\pi_\theta(s, a) ds.$$
The Actor-Critic approach:

- **Critic**: estimates parameters of the reward function $E(R|s, a)$
- **Actor**: updates the parameters indexing the parametrized policy, using the estimates provided by the critic, i.e. finding $\theta$ that maximizes

$$V^*(\theta) = \int_{s \in S} d(s) \sum_{a \in A} E(R|s, a)\pi^\theta(s, a)ds.$$ 

- The Actor-Critic approach uses two models for the mean of the reward, $E(R|s, a)$, (critic) and the policy, $\pi^\theta$ (actor).
Theoretical results from CS/Stats literature focus on regret bounds but we need inferential tools to resolve practical issues in mobile health e.g., to determine which variables are important.

Lei, Tewari and Murphy (2017) established consistency and asymptotic normality of actor-critic bandit. Using these results, test statistics could be constructed to test whether variables of interest are important.

As in the usual statistical procedures based on the assumed models, such constructed test statistics are not robust with respect to model misspecification.

The problem of maximizing the average reward, $V^*$ lends itself to a different problem from usual hypothesis testing.

We propose how to construct a robust test for usefulness of variables in sequential decision making.
Actor-Critic Bandit

- Expected reward: $E(R(t)|S = s, A = a) = r(s, a)$.
- Model for the expected reward: $r(s, a) = f(s, a)^T \mu$
- Model for the policy (probability of allocation): e.g.,

$$\pi_\theta(s, a) = \frac{\exp(\theta_0 a + \theta_1 g_1(s_1) + \theta_2 g_2(s_2))}{\sum_{a \in A} \exp(\theta_0 a + \theta_1 g_1(s_1) + \theta_2 g_2(s_2))}$$

- We have an optimization problem: Find $\theta$ that maximizes the marginal expected reward $V^*(\theta)$ using the estimate for $r(s, a)$ (Critic), where

$$V^*(\theta) = \int_{s \in S} d(s) \sum_{a \in A} \hat{r}(s, a) \pi_\theta(s, a) ds.$$
Algorithm 1: Actor-Critic Bandit

Critic initialization: $B(0) = \xi I_{d \times d}$; $A(0) = 0_{k \times 1}$;
Actor initialization: $\theta_0$;
Start from $t = 0$.

while $t \leq T$ do

At time $t$, observe context $s_t$;
Draw an action $a(t)$ according to $\pi_{\hat{\theta}_{t-1}}(s_t, a)$;
Observe reward $R_t$;
Critic update: $B(t) = B(t - 1) + f(s_t, a_t)f(s_t, a_t)^T$,
$A(t) = A(t - 1) + f(s_t, a_t)R_t$, $\hat{\mu} = B(t)^{-1}A(t)$
Actor update: $\hat{\theta}_t = \text{argmax}_\theta \sum_{t=1}^{T} \hat{r}(s_t, a)\pi_\theta(s_t, a)$
Go to time $t + 1$

end while
Assume that $s_i$ are iid

Truncate $\hat{r}(s, a)$ to enforce boundedness.

Instead of maximizing the marginal mean $V^*(\theta)$, maximize penalized version

$$V^*(\theta) - \lambda L(\theta) = V^*(\theta) - \lambda \theta^T E[g(S)^T g(S)] \theta$$

so that $\theta^*$, the value that $\hat{\theta}$ converges to lies in a bounded set.

We use a simpler regularized term and maximize

$$J^*_\lambda(\theta) = V^*(\theta) - \lambda \theta^T \theta$$

The proof goes through with this modified loss function
Asymptotic results of Actor-Critic Bandit by Lei et al. (2017)

- Asymptotic properties of the critic:
  \[ \sqrt{t}(\hat{\mu}_t - \mu^*) \sim N(0, \sigma^2[E(f(S, A)f(S, A)^T)]^{-1}) \]

- Asymptotic properties of the actor:
  \[ \sqrt{t}(\hat{\theta}_t - \theta^*) \sim N(0, [J_{\theta\theta}(\mu^*, \theta^*)]^{-1} \nabla [J_{\theta\theta}(\mu^*, \theta^*)]^{-1}) \]

- Note that \( \hat{\mu}_t \) is not related to \( \hat{\theta}_t \), but \( \hat{\theta}_t \) is related to \( \hat{\mu}_t \).
Lei et al. (2017) showed the asymptotic normality of the estimated optimal policy

- This enables hypothesis testing of parameters underlying the actor critic.

- To conduct tests of whether certain variables are needed for the policy, we need to use the properties of the actor. Some variables may affect the reward, but may not affect the policy.

- Now we have a framework whereby we can conduct such tests, and address hypotheses about the policy.
When the policy model is correctly specified:

- Note that \( V^*(\theta) \) is a function of \( \pi \), and we would like to find \( \pi \) that maximizes \( V^*(\theta) \),

\[
V^*(\theta) = \int_{s \in S} d(s) \sum_{a \in A} r(s, a) \pi_\theta(s, a) \, ds
\]

- A sample version of \( V^*(\theta) \) is

\[
\tilde{V}^*(\theta) = T^{-1} \sum_{t=1}^{T} \sum_{a_t \in A} [\hat{r}(s_t, a_t) \hat{\pi}_\theta(s_t, a_t)]
\]

- A test statistic for \( \theta \) can be derived by using the score-like quantity i.e. \( \frac{\partial}{\partial \theta} \tilde{V}^*(\theta) \)
Testing usefulness of watch variable

The validity of statistical tests is predicated on correctly specified models

- Note that there are two models in $V^*$: $r(s, a)$ and $\pi_{\theta}$:

\[
V^*(\theta) = \int_{s \in S} d(s) \sum_{a \in A} r(s, a)\pi_{\theta}(s, a)ds
\]

- There are two sources of model misspecification: the critic ($r(s, a)$) and actor ($\pi_{\theta}$). What happens to the validity of the test when one of the models is misspecified?

- There are common tricks used in policy gradient methods that provide guidance on modeling $r(s, a)$. One of which (compatibility condition) addresses the issue of misspecifying $r(s, a)$.

- Then we present the main result on how to deal with misspecification of $\pi_{\theta}$.
Commonly used tricks in policy gradient methods: use of Advantage

- Note that $V^*(\theta)$ is a function of $\pi$, and we would like to find $\pi$ that maximizes $V^*(\theta)$,

$$V^*(\theta) = \int_{s \in S} d(s) \sum_{a \in A} r(s, a) \pi_{\theta}(s, a) ds$$

- It is common to replace reward, $r(s, a)$, with residual or Advantage, $r(s, a) - h(s)$, because replacing does not affect the maximum,

$$\frac{\partial}{\partial \theta} \int_{s \in S} d(s) \sum_{a \in A} h(s) \pi_{\theta}(s, a) ds = \int_{s \in S} d(s) h(s) \frac{\partial}{\partial \theta} \sum_{a \in A} \pi_{\theta}(s, a) ds = 0.$$ 

By choosing $h(s) = E_a\{r(s, a)\}$, $r(s, a)$ can be replace with $[r(s, a) - E_a\{r(s, a)\}]$:

$$\frac{\partial}{\partial \theta} V^*(\theta) = \int_{s \in S} d(s) \sum_{a \in A} [r(s, a) - E_a\{r(s, a)\}] \pi_{\theta}(s, a) ds$$
Commonly used tricks in policy gradient methods: use of compatibility condition

Compatibility condition: \( \frac{\partial}{\partial \theta} \log \pi_\theta(s, a) = \frac{\partial}{\partial w} m_w(s, a). \)

- For the Critic, minimize the expected sum of squares by solving
  \[
  \int_{s \in S} d(s) \sum_{a \in A} \{ r(s, a) - m_w(s, a) \} \frac{\partial}{\partial w} m_w(s, a) \pi_\theta(s, a) ds = 0
  \]

- For the Actor, maximize \( V^*(\theta) \) by solving
  \[
  \frac{\partial}{\partial \theta} V^*(\theta) = \int_{s \in S} d(s) \sum_{a \in A} r(s, a) \frac{\partial}{\partial \theta} \pi_\theta(s, a) ds = 0
  \]

- By the compatibility condition:
  \[
  \int_{s \in S} d(s) \sum_{a \in A} r(s, a) \frac{\partial}{\partial \theta} \pi_\theta(s, a) ds = \int_{s \in S} d(s) \sum_{a \in A} m_w(s, a) \frac{\partial}{\partial \theta} \pi_\theta(s, a) ds
  \]

- When the model in the critic is misspecified, we can find the optimal policy if the critic is modeled to satisfy the compatibility condition. We will focus on the case that optimal policy is not in the function
Robustness in the statistics literature

- Rosenblum and van der Laan (2009) presented a class of models that was robust to misspecification;
  - Linear models, GLM (Poisson, binomial)
  - Survival models, Cox model (Kim 2013)

- Robustness means: “We say a hypothesis test at level $\alpha$ is asymptotically robust to misspecification if for any data generating distribution satisfying the null hypothesis [of no treatment effect], the asymptotic probability of rejecting the null hypothesis is at most $\alpha$. ” (Rosenblum and van der Laan 2009)

- Key assumption: orthogonality of the treatment assignment variable to other variables

- Hypothesis: testing treatment effect within strata of baseline variables
Key idea of robust test in regression settings

- $H_0: E(Y|X,Z) = E(Y|Z)$. Let the misspecified model $m$ be

$$m(\beta) = \beta_0 + \beta_1 X + \beta_2 Z$$

Rewrite $\beta = (\beta_1, \bar{\beta})$, where $\bar{\beta} = (\beta_0, \beta_2)$. The score-like function is

$$S(\beta_1, \bar{\beta}) = \begin{pmatrix} S_0(\beta_1, \bar{\beta}) \\ S_1(\beta_1, \bar{\beta}) \\ S_2(\beta_1, \bar{\beta}) \end{pmatrix} = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ X_i \\ Z_i \end{pmatrix} (Y_i - m_i(\beta_1, \bar{\beta}))$$

- Under the null, $\beta_1 = 0$. To use $S_1(0, \bar{\beta})$ as a test statistic, we need to replace $\bar{\beta}$ with an estimate.
- Let $\tilde{\beta}$ be the solution of $\begin{pmatrix} S_0(0, \bar{\beta}) \\ S_2(0, \bar{\beta}) \end{pmatrix} = 0$, and $\beta^*$ be the limit of $\tilde{\beta}$.
- We have $E(S_0(0, \beta^*)) = 0$ and $E(S_2(0, \beta^*)) = 0$. 
Key idea of robust test in regression settings

- We have the unbiasedness of $S_0$ and $S_2$ under the null.
- For $S_1(0, \tilde{\beta})$ to be a valid test statistic, a key is to show
  \[ E(S_1(0, \beta^*)) = 0 \]
  under the null $H_0: E(Y|X,Z) = E(Y|Z)$.
- Using the fact $E(S_0(0, \beta^*)) = 0$, when $X$ and $Z$ are independent,
  \[ E(S_1(0, \beta^*)) = \sum_{i=1}^{n} E(X_i)E(E(Y_i|Z_i) - m_i(0, \bar{\beta})) = E(X)E(S_0(0, \beta^*)) = 0. \]
- This trick works by (i) expressing $S_1$ as a product of a function of $X$ and that of $Z$, and (ii) applying the unbiasedness of $S_0$. 
Robust test for actor-critic bandit

- Let $s = (s_1, s_2)$. $H_0$: $\pi^*(s, a) = \pi^*(s_2, a)$. Does the optimal policy depend on $s_1$?
- When $\pi(s, a)$ does not belong to the function class $\pi_\theta(s, a)$, what happens to the hypothesis test based on the maximizer of

$$\tilde{V}(\theta) = T^{-1} \sum_{t=1}^{T} \sum_{a_t \in A} [\hat{m}(s_t, a_t)\pi_\theta(s_t, a_t)]$$

where $m(s, a)$ is the misspecified model.

- Consider an augmented model

$$\pi_\theta(s, a) = \frac{\exp(\eta_a)}{\sum_{a \in A} \exp(\eta_a)} = \frac{\exp(\theta_0a + \theta_1g_1(s_1) + \theta_2g_2(s_2) + \theta_3g^*(s_2))}{\sum_{a \in A} \exp(\theta_0a + \theta_1g_1(s_1) + \theta_2g_2(s_2) + \theta_3g^*(s_2))}$$

where $g^*(s_{2t}) = E(g_1(s_{1t})|s_{2t})$.

- Without loss of generality, let $\theta = (\theta_1, \bar{\theta})$, where $\bar{\theta} = (\theta_0, \theta_2, \theta_3)$. 
Robust test for actor-critic bandit

We have a set of score-type of equations.

- Let $U(\theta) := \frac{\partial}{\partial \theta} \tilde{V}(\theta)$.
- Under the null, $\theta_1 = 0$. To use $U_1(0, \bar{\theta})$ as a test statistic, we need to replace $\bar{\theta}$ with $\tilde{\theta}$, which is the solution of

$$
\begin{pmatrix}
U_0(0, \tilde{\theta}) \\
U_2(0, \tilde{\theta}) \\
U_3(0, \tilde{\theta})
\end{pmatrix} = \sum_{t=1}^{T} \sum_{a_t \in A} \begin{pmatrix}
1 \\
\ddot{g}_2(s_{2t}) \\
\ddot{g}^*(s_{2t})
\end{pmatrix} [\hat{m}(s_t, a_t) \nabla_{\eta} \pi_{\theta}(s_{2t}, a_t)] = 0
$$

- Let $\theta^*$ be the limit of $\tilde{\theta}$.
- We have $E(U_0(0, \theta^*)) = 0$, $E(U_2(0, \theta^*)) = 0$ and $E(U_3(0, \theta^*)) = 0$. 
We have a set of score-type of equations.

- For $U_1(0, \tilde{\theta})$ to be a valid test statistic for $H_0: \pi^*(s, a) = \pi^*(s_2, a)$, a key is to show that $EU_1(0, \theta^*) = 0$, where

$$U_1(0, \theta) = \sum_{t=1}^{T} \sum_{a_t \in A} g_1(s_{1t}) [\hat{m}(s_t, a_t) \nabla_{\eta} \pi_{\theta}(s_{2t}, a_t)]$$

- Unlike the regression setting,

$$EU_1(0, \theta) = E \sum_{t, a_t} g_1(s_{1t}) [\hat{m}(s_t, a_t) \nabla_{\eta} \pi_{\theta}(s_{2t}, a_t)] \neq E(g_1(s_1)) EU_0(0, \theta)$$

since $U_0(0, \bar{\theta})$ is a function of $(s_1, s_2)$ due to $m(s_t, a_t)$. 
The components of $U_1(\theta, 0)$ cannot be easily separated. To address this, we consider centering the covariate in the reward model:

- We need to model $m(s_t, a_t)$ so that $U_0(0, \bar{\theta})$ is a function of $s_2$ only.
- For a model

$$m((s_1, s_2), a) = \mu_0 + \mu_1 f_1(s_1, a) + \mu_2 f_2(s_2, a)$$

to satisfy this condition, we can replace $f_1(s_1, a)$ with $f_1(s_1, a) - \bar{f}_1(s_1)$ where

$$\bar{f}_1(s_1) = \frac{\sum_{a \in A} f_1(s_1, a) \nabla_{\eta \pi_{\theta^*}}(s_2, a)}{\sum_{a \in A} \nabla_{\eta \pi_{\theta^*}}(s_2, a)}$$

We can verify

$$\sum_{a_t \in A} \{f_1(s_{1t}, a_t) - \bar{f}_1(s_{1t})\} \nabla_{\eta \pi_{\theta^*}}(s_{2t}, a_t) = 0.$$
Now with \( m((s_1, s_2), a) = \mu_0 + \mu_1(f_1(s_1, a) - \bar{f}_1(s_1)) + \mu_2 f_2(s_2, a) \), let

\[
b_\theta(s_{2t}) := \sum_{a_t \in \mathcal{A}} m(s_t, a_t) \nabla_{\theta} \pi_{\theta^*}(s_{2t}, a_t)
\]

Then \( U_0(0, \bar{\theta}) = \sum_{t=1}^{T} b_\theta(s_{2t}) \), due to centering the covariate.

Recall we have the unbiasedness of \( U_0 \) at \( \theta^* \), \( EU_0(0, \theta^*) = 0 \).

If \( s_1 \) and \( s_2 \) are independent,

\[
EU_1(0, \theta^*) = E \sum_{t=1}^{T} g(s_{1t}) b_{\theta^*}(s_{2t}) = 0
\]
If $s_1$ and $s_2$ are not independent,

$$EU_1(0, \theta^*) = E\sum_{t=1}^{T} E(g(s_{1t})|s_{2t})b_{\theta^*}(s_{2t}) = E\sum_{t=1}^{T} g^*(s_{2t})b_{\theta^*}(s_{2t}) = 0,$$

since $EU_3(0, \theta^*) = E\sum_{t=1}^{T} g^*(s_{2t})b_{\theta^*}(s_{2t}) = 0$.

By centering $f_1(s_1, a)$ in the reward model and augmenting $g^*(s_{2t})$ for $\pi$ model, we can make $U_1(0, \tilde{\theta})$ a valid statistic.
Handling nuisance parameters

- Let the empirical version of $b_\theta(s_{2t})$ be
  $$\hat{b}_\theta(s_{2t}) := \sum_{a_t \in A} \hat{m}(s_t, a_t) \nabla_{\theta} \pi^\star(s_{2t}, a_t).$$

  $$\hat{U}_1(0, \tilde{\theta}) = T^{-1} \sum_{t=1}^{T} g_1(s_{1t}) \hat{b}_{\tilde{\theta}}(s_{2t})$$

- The test statistic is
  $$\hat{U}_1(0, \tilde{\theta})^T V(\hat{U}_1(0, \tilde{\theta}))^{-1} \hat{U}_1(0, \tilde{\theta}).$$

  The variance of $\hat{U}_1(0, \tilde{\theta})$, $V(\hat{U}_1(0, \tilde{\theta}))$, can be obtained using the usual linearization method.
The problem of testing the utility of passive data collected by wearable devices is a real issue, due to the cost of collecting data.

We considered testing variables in the actor critic bandit, but considered inference when models used in the algorithms are not correctly specified.

We can construct robust tests for the policy parameter, with an augmented model.

This work illustrates that due to model assumptions, inference with the bandit inherits problems such as misspecification; however, existing tools in the literature do not apply, the unique structure of the objective function in this case calls for special treatment.
Future work

- We assume that hypothesis tests are conducted after a large number of trials $T$ at one point. Sequential or repeated hypothesis testing may be an alternative.

- Tewari and Murphy addressed issues that need consideration when using bandits in mHealth. Among these issues are variable selection (to address data traffic) and the selection of initial values.

- To obtain reasonable initial values, offline policy evaluation is important. Efforts to share the data or to build a large database may be needed.
A Note on Using Regression Models to Analyze Randomized Trials: Asymptotically Valid Hypothesis Tests Despite Incorrectly Specified Models.
*Biometrics, 69*(1), 282-289.

Lei, H., Tewari, A., Murphy, S. (2017).
An actor-critic contextual bandit algorithm for personalized interventions using mobile devices.
*arXiv: 1706.09090*

Using regression models to analyze randomized trials: Asymptotically valid hypothesis tests despite incorrectly specified models.
*Biometrics, 65*(3), 937-945.

Development of a smartphone application for eating disorder self-monitoring.
Thank you!