Robust Tests in Online Decision-Making

Gi-Soo Kim[1], **Jane P. Kim**[2], **Hyun-Joon Yang**[2]

[1]Department of Industrial Engineering & Artificial Intelligence Graduate School, UNIST
[2]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine
gisookim@unist.ac.kr, janepkim@stanford.edu, yanghyun@stanford.edu

Abstract

Bandit algorithms are widely used in sequential decision problems to maximize the cumulative reward. One potential application is mobile health, where the goal is to promote the user's health through personalized interventions based on user specific information acquired through wearable devices. Important considerations include the type of, and frequency with which data is collected (e.g. GPS, or continuous monitoring), as such factors can severely impact app performance and users' adherence. In order to balance the need to collect data that is useful with the constraint of impacting app performance, one needs to be able to assess the usefulness of variables. Bandit feedback data are sequentially correlated, so traditional testing procedures developed for independent data cannot apply. Recently, a statistical testing procedure was developed for the actor-critic bandit algorithm. An actor-critic algorithm maintains two separate models, one for the actor, the action selection policy, and the other for the critic, the reward model. The performance of the algorithm as well as the validity of the test are guaranteed only when the critic model is correctly specified. However, misspecification is frequent in practice due to incorrect functional form or missing covariates. In this work, we propose a modified actor-critic algorithm which is robust to critic misspecification and derive a novel testing procedure for the actor parameters in this case.

Introduction

Bandit algorithms apply to sequential decision problems. We assume a set of candidate actions, or *arms*, is revealed sequentially to a learning agent along with side information called contexts. The agent can pull one arm at a time and receives a corresponding reward. The expected value of the reward is an unknown function of the context information of the chosen action. The goal of the agent is to adaptively learn an action allocation policy so as to achieve high cumulative reward. The main challenge is the exploration-exploitation trade-off, which represents the dilemma between pulling arms that the agent is uncertain about for the sake of learning (exploration) and pulling the best arm based on current, limited knowledge (exploitation).

Bandit algorithms can be particularly useful in the context of personalizing health interventions in Mobile Health (Tewari and Murphy 2017). The goal of Mobile Health (mHealth) apps is to promote the user's health through personalized interventions tailored to the user specific information acquired through devices such as phones or wearable devices. One important issue related to mHealth apps is that the frequency of data queries (e.g. queries to the Health Kit API) impacts the app performance. As queries become increasingly frequent, the processing time slows down the app from the user perspective, which can result in low adherence to the app and hence low reward. Hence, the frequency of data queries and the reward feedback go hand in hand. Beyond performance related costs, there could also be costs of ethical valence (e.g. privacy) associated with querying data and thus it is important to collect only data that are correlated with the reward. Currently, there is little work on assessing the utility of variables collected by wearables.

We consider testing the utility of context variables for an actor-critic bandit algorithm (Lei, Tewari, and Murphy 2017). An actor-critic bandit algorithm maintains two separate parameterized models, one for the actor, the action allocation policy, and the other for the critic, the reward model. The focus of this work is on testing the variables used in the actor model, which requires that asymptotic distributions of the actor parameter estimates are known. Lei, Tewari, and Murphy (2017) proved that when the reward model is linear and is correctly specified by the critic, then the actor parameter estimates converge in probability to the parameters of the optimal policy and asymptotically follow a normal distribution.

Based on the asymptotic normality of the actor parameter estimates, we can apply a Z-test to assess the significance of the actor parameters. The validity of model-based tests, such as those of Lei, Tewari, and Murphy (2017), relies on the assumption that the linear model is correctly specified; in other words, that the assumed statistical model represents the true reward function. Linear functions may, however, fail to accurately represent the true nature of the reward function, and often there is no a priori reason to hypothesize the reward should be of a certain functional form. When the parameterized critic model is not correctly specified (i.e. the true reward model is of a different form than the working model), asymptotic normality may not hold. In this paper, we propose a new actor-critic algorithm and a testing procedure that is robust to the misspecification of the critic

model. The main contributions of our paper are as follows:

- We propose a new actor-critic algorithm where the actor parameter estimates converge to the parameters of the optimal policy even when the reward model is misspecified by the critic.
- We show that in the new algorithm, critic and actor parameter estimates asymptotically follow a normal distribution.
- We conduct experiments on synthetic data and real data and show that our testing procedure appropriately assess the significance of the parameters.

Related Works on Robust Bandits

Our work is distinct from the limited literature on robust bandits. First, the focus of the existing works (Tang et al. 2021; Hao et al. 2019; Zhu et al. 2018) is on the robustness to the noise of the rewards. For example, Tang et al. (2021) and Hao et al. (2019) developed Upper-Confidence Bound (UCB) algorithms without requiring to specify the tail property of the reward, and Zhu et al. (2018) developed an actor-critic (AC) algorithm that is robust to outliers. These methods however, are built upon the linearity of the reward model. Ghosh, Chowdhury, and Gopalan (2017) developed a new algorithm that maintains the sublinear regret in a model with large deviations from linearity, but is restricted to the case where the action feature set is fixed over time. Hence even under large deviations, the problem can still be addressed by a multi-armed bandit algorithm with regret scaling with the number of arms instead of feature dimension.

Second, we consider the AC algorithm, which, apart from ϵ-greedy, is unique in that asymptotic distributions of the parameter estimates are known; there are currently no established statistical testing procedures for the UCB or Thompson sampling algorithms. We consider the impact of misspecification on the validity of inferential testing of the utility of contextual variables (i.e. significance of actor parameters in the AC algorithm). To the best of our knowledge, no other work addresses the robustness of inferential testing in the context of the actor-critic algorithm. In a related work on the ϵ-greedy bandit, Chen, Lu, and Song (2020) derived the asymptotic properties of a weighted least squares estimator (LSE) for misspecified reward models. The authors demonstrated that the weighted LSE has asymptotic normal distribution with the mean being the least false linear parameter. While both this and our approach offer robust tests, ours offers a directed approach to exploration, which may be efficient and desirable in the mobile health setting where the action space is large.

Preliminaries

Problem Formulation

We first formulate the bandit problem. At each time t, the learning agent can pull one arm among N alternative arms. The i-th arm ($i = 1, \cdots, N$) returns a random reward $r_{t,i}$ with unknown mean when it is pulled. Prior to arm selection, a finite-dimensional context vector $b_{t,i} \in \mathbb{R}^d$ for each arm i is revealed to the agent. The agent tailors his(her) choice based on this contextual information. Let a_t denote the arm index pulled at time t. Then the goal of the agent is to maximize the cumulative sum of the rewards r_{t,a_t} over a finite time horizon T. We assume that the full set of contexts $b_t = \{b_{t,1}, \cdots, b_{t,N}\}$ and the full set of rewards $r_t = \{r_{t,1}, \cdots, r_{t,N}\}$ are independently and identically (i.i.d.) distributed over time. Also, without loss of generality, we assume the L_2-norm of $b_{t,i}$ is bounded by 1, i.e., $||b_{t,i}||_2 \leq 1$.

Notations

We denote the L_2-norm of a vector x as $||x||_2$, the set of natural numbers from 1 to N as $[N]$, the set of all natural numbers as $\mathbb{N}$, the d-dimensional identity matrix as $I^{d\times d}$, and the d-dimensional vector with all elements equal to 0 as 0_d.

Actor-Critic Bandit Algorithm for Linear Rewards

Under linear reward assumption $\mathbb{E}[r_{t,i}|b_{t,i}] = b_{t,i}^T\mu^*$ for some $\mu^* \in \mathbb{R}^d$ with $||\mu^*||_2 \leq 1$, Lei, Tewari, and Murphy (2017) proposed the actor-critic bandit algorithm (Algorithm 1) which learns two parametrized models, the critic and the actor. The critic for the i-th arm is a linear function of the i-th context variable with parameter $\mu \in \mathbb{R}^d$, $b_{t,i}^T\mu$. The actor is the action allocation probability and is parametrized by a softmax function with parameter $\theta \in \mathbb{R}^d$, i.e., the probability of pulling the i-th arm at time t is $\pi_\theta(b_t, i) = \exp(b_{t,i}^T\theta)/\{\sum_{j=1}^N \exp(b_{t,j}^T\theta)\}$. Lei, Tewari, and Murphy (2017) define the optimal parameter θ^* as the value that maximizes the penalized expected reward,

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}\left[\sum_{i=1}^N b_{t,i}^T\mu\ \pi_\theta(b_t, i)\right] - \lambda\theta^T\theta,$$

where $\lambda > 0$ and the expectation is taken over the distribution of b_t. The penalty term $-\lambda\theta^T\theta$[1] is introduced to constrain the norm of θ. Due to the penalty, there exists $\gamma > 0$ such that $\gamma < \pi_{\theta^*}(b_t, i) < 1 - \gamma$ for every i with high probability. This guarantees treatment variety, which prevents habituation and increases user engagement in many applications including mHealth. Also, when the expected rewards of the arms are the same so that the $\mathbb{E}\left[\sum_{i=1}^N b_{t,i}^T\mu\ \pi_\theta(b_t, i)\right]$ term does not change according to the values of θ, θ^* is unique at $\theta^* = 0_d$.

The estimator $\hat{\mu}$ of the critic parameter μ is the Ridge estimator using the context and reward pair of the chosen arms. The estimator $\hat{\theta}$ of the policy parameter θ is the maximizer of the estimate of the penalized expected reward, $\frac{1}{t}\sum_{\tau=1}^t \sum_{i=1}^N \hat{r}_{\tau,i}\pi_\theta(b_\tau, i) - \lambda\theta^T\theta$, where $\hat{r}_{\tau,i}$ is the truncated estimate of the reward defined in Algorithm 1, line 5. When the true critic parameter μ^* has $||\mu^*||_2 \leq 1$ and $\hat{\mu}_t$ converges to μ^*, the truncated reward estimate $\hat{r}_{t,i}$ approaches the untruncated estimate, $b_{t,i}^T\hat{\mu}_t$.

[1]The presented penalty form is a special case of the penalty proposed in Lei, Tewari, and Murphy (2017).

Algorithm 1: Actor-Critic algorithm for linear reward [Lei et al.,2017]

1: Set $B = \xi I^{d\times d}$, $y = 0_d$, $\lambda > 0$, $\xi > 0$.
2: **for** $t = 1, \cdots, T$ **do**
3: Pull arm a_t according to probability $\left\{\pi_{\hat{\theta}_{t-1}}(b_t, i)\right\}_{i=1}^{N}$ and get reward r_{t,a_t}.
4: **Critic update**:

$$B \leftarrow B + b_{t,a_t} b_{t,a_t}^T, \;\; y \leftarrow y + b_{t,a_t} r_{t,a_t}, \;\; \hat{\mu}_t \leftarrow B^{-1} y.$$

5: $\hat{r}_{\tau,i} \leftarrow \max(-2, \min(2, b_{\tau,i}^T \hat{\mu}_t))$ for $i \in [N]$, $\tau \in [t]$.
6: **Actor update**:

$$\hat{\theta}_t \leftarrow \underset{\theta}{\operatorname{argmax}} \frac{1}{t} \sum_{\tau=1}^{t} \sum_{i=1}^{N} \hat{r}_{\tau,i} \pi_\theta(b_\tau, i) - \lambda \theta^T \theta.$$

7: **end for**

The boundedness of $\hat{r}_{\tau,i}$ and the penalty term ensures that $\hat{\theta}_t$ is bounded. This guarantees that there exists $\gamma > 0$ such that $\gamma < \pi_{\hat{\theta}_t}(b_t, i) < 1 - \gamma$ for every i. This prevents $\pi_{\hat{\theta}_t}(b_t, i)$ from concentrating on a single arm and induces a degree of exploration.

Lei, Tewari, and Murphy (2017) showed that under some regular assumptions on the distribution of the contexts and rewards, $\hat{\mu}_t$ and $\hat{\theta}_t$ converge in probability to μ^* and θ^* respectively and are asymptotically normally distributed, hereby enabling a testing procedure.

Misspecification of Models

The validity of model-based testing is predicated on correctly specified models. However, misspecification is frequent in practice due to incorrect functional forms or missing covariates. In the statistics literature, robustness has been considered in the context of using models to test causal effects from data collected in experiments. Linear and GLM regression models (Rosenblum and Van Der Laan 2009) and proportional and multiplicative hazards models (Kim 2013) have been shown to be robust to misspecification when considering the test of the coefficient of the treatment assignment in the context of randomized trials. However in bandit settings, asymptotic normality is not guaranteed to hold when the working model is incorrect. In this paper, we consider the case where the critic is misspecified.

Inference from Bandit Feedback Data

Besides Lei, Tewari, and Murphy (2017), there is a recent growing body of literature on deriving the distribution of the parameter estimates from bandit feedback data. Bandit feedback data are not i.i.d. but are correlated due to adaptivity. This causes complexity in deriving the distribution of the estimates. Zhang, Janson, and Murphy (2021) recently showed the asymptotic distribution of M-estimators from bandit data. This work considered correctly specified reward models only. Chen, Lu, and Song (2020) derived the asymptotic normality of the ordinary and weighted least-squares estimators when data are accumulated by a ε-greedy algorithm, in both cases where the reward model is linear or not linear. When the reward model is not linear, they proved that the estimator with inverse-probability weighting converges to the normal distribution with mean being the least false parameter in terms of the population distribution of the contexts. Since the action decision in ϵ-greedy algorithms is based on reward estimate values, a robust test on the utility of the variables could be conducted by testing the significance of the least false parameters. However, as aforementioned earlier, we note that ε-greedy performs uniform exploration over context spaces which may be undesirable when the action space is large.

Compatibility Condition in Actor-Critic Algorithm

The algorithm of Lei, Tewari, and Murphy (2017) and the theoretical derivation therein exploit the fact that the true reward model is linear. The true nature of the reward can however be far from linear. Sutton et al. (2000) proved the following Lemma 1 which implies that if the reward model and policy model are both differentiable with respect to their parameters and satisfy the compatibility condition, the algorithm converges to the optimal policy $\pi_{\theta*}$ though the critic model may be misspecified. If we denote the critic model parameterized by μ as $m_\mu(\cdot)$, the compatibility condition states:

$$\dot{\pi}_\theta(b_t, i)/\pi_\theta(b_t, i) = \dot{m}_\mu(b_{t,i}), \tag{1}$$

where $\dot{\pi}_\theta(b_t, i) = \frac{\partial}{\partial\theta}\pi_\theta(b_t, i)$ and $\dot{m}_\mu(b_{t,i}) = \frac{\partial}{\partial\mu} m_\mu(b_{t,i})$.

Lemma 1. *(Theorem 2 of Sutton et al. (2000)) Let*

$$J(\theta) = \mathbb{E}_{b,r}\left[\sum_{i=1}^{N} r_{t,i}\pi_\theta(b_t, i)\right] - \lambda\theta^T\theta, \tag{2}$$

where $\mathbb{E}_{b,r}(\cdot)$ denotes the expectation over both the context and reward. Suppose the critic parameter μ minimizes

$$U(\mu, \theta) := \mathbb{E}_{b,r}\left[\sum_{i=1}^{N} \left\{r_{t,i} - m_\mu(b_{t,i})\right\}^2 \pi_\theta(b_t, i)\right],$$

and the actor parameter θ maximizes

$$J(\mu, \theta) := \mathbb{E}_b\left[\sum_{i=1}^{N} m_\mu(b_{t,i})\, \pi_\theta(b_t, i)\right] - \lambda\theta^T\theta.$$

Then if $\pi_\theta(\cdot)$ and $m_\mu(\cdot)$ satisfy the compatibility condition (1), the actor parameter θ satisfies $\frac{\partial}{\partial\theta}J(\theta) = 0$.

Sketch of Proof. The parameters μ and θ jointly solve $U_\mu(\mu, \theta) = 0$ and $J_\theta(\mu, \theta) = 0$, where $U_\mu(\mu, \theta) = -\frac{1}{2}\frac{\partial}{\partial\mu}U(\mu, \theta)$ and $J_\theta(\mu, \theta) = \frac{\partial}{\partial\theta}J(\mu, \theta)$. We have,

$$U_\mu(\mu, \theta) = \mathbb{E}_{b,r}\left[\sum_{i=1}^{N} \left\{r_{t,i} - m_\mu(b_{t,i})\right\} \dot{m}_\mu(b_{t,i})\pi_\theta(b_t, i)\right] \tag{3}$$

and

$$J_\theta(\mu,\theta) = \mathbb{E}_b\left[\sum_{i=1}^{N} m_\mu(b_{t,i})\,\dot{\pi}_\theta(b_t,i)\right] - 2\lambda\theta \quad (4)$$

Due to (1) and the facts that (3)= 0, and (4)= 0, the parameter θ satisfies

$$\frac{\partial}{\partial\theta}J(\theta) = \mathbb{E}_{b,r}\left[\sum_{i=1}^{N} r_{t,i}\,\dot{\pi}_\theta(b_t,i)\right] - 2\lambda\theta = 0.$$

Note that in Lemma 1, $J(\theta)$ is defined in terms of the true rewards $r_{t,i}$'s, while $J(\mu,\theta)$ replaces them with $m_\mu(b_t,i)$'s. If the true reward model is linear, i.e., if $\mathbb{E}[r_{t,i}|b_{t,i}] = b_{t,i}^T\mu^*$, and if $m_\mu(b_t,i) = b_{t,i}^T\mu$, then we have $J(\theta) = J(\mu^*,\theta)$. However when the true model is not linear, $J(\theta)$ and $J(\mu,\theta)$ are completely different functions. Sutton et al. (2000) show that (1) is satisfied when the actor model is a softmax function and the critic model is linear in the same context vectors as the policy, except they should be centered to have weighted mean 0:

$$\textbf{critic}: \quad m_{\mu,\theta}(b_{t,i}) = \left\{ b_{t,i} - \sum_{j=1}^{N}\pi_\theta(b_t,j)b_{t,j}\right\}^T \mu \quad (5)$$

$$\textbf{actor}: \quad \pi_\theta(b_t,i) = \frac{\exp\left(b_{t,i}^T\theta\right)}{\sum_{j=1}^{N}\exp\left(b_{t,j}^T\theta\right)} \quad (6)$$

The model (5) can be viewed as the approximation of the *advantage* function (Baird III 1993). The advantage function enables to discard variables that do not vary by arm (e.g., age of the user). We would still need such variables if we model the reward instead of the advantage. From now on we denote the model (5) as $m_{\mu,\theta}(\cdot)$ instead of $m_\mu(\cdot)$ to show its dependency on θ as well. Meanwhile, since $\sum_{i=1}^{N}\dot{\pi}_\theta(b_t,i) = 0$, equation (4) is equivalent to

$$J_\theta(\mu,\theta) = \mathbb{E}_b\left[\sum_{i=1}^{N} b_{t,i}^T\mu\,\dot{\pi}_\theta(b_t,i)\right] - 2\lambda\theta.$$

So we redefine

$$J(\mu,\theta) = \mathbb{E}_b\left[\sum_{i=1}^{N} b_{t,i}^T\mu\,\pi_\theta(b_t,i)\right] - \lambda\theta^T\theta. \quad (7)$$

We can find the value of θ satisfying $(4) = 0$ as the maximizer of the redefined $J(\mu,\theta)$.

Proposed Algorithm

We propose a new actor-critic algorithm which uses (5) and (6) to model the reward and action selection policy. We consider the case where the true functional form of the reward model may not be linear. In this case, we re-define the target parameters μ^* and θ^* as

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, J(\theta), \quad \mu^* = \underset{\mu}{\operatorname{argmin}}\, U(\mu,\theta^*),$$

where $J(\theta)$ is defined in (2) and

$$U(\mu,\theta) = \mathbb{E}_{b,r}\left[\sum_{i=1}^{N}\left\{r_{t,i} - m_{\mu,\theta}(b_{t,i})\right\}^2\pi_\theta(b_t,i)\right]. \quad (8)$$

Under (5) and (6) which satisfy the compatibility condition, $\theta^* = \operatorname{argmax}_\theta J(\mu^*,\theta)$, where $J(\mu,\theta)$ is redefined in (7).

We assume that the arguments that achieve the maximum(argmax) and minimum(argmin) both exist in the parameter space that we consider. While the definition of θ^* is the same as the original definition, we notice that the definition of μ^* now depends on the value of θ^*.

Estimating Functions for μ^* and θ^*

The target parameters μ^* and θ^* are the values that jointly minimize (8) with respect to μ and maximize (7) with respect to θ. To consistently estimate the parameters, we use as estimating functions the empirical versions of (8) and (7) that converge in probability to (8) and (7) respectively. Suppose we use the residual mean square (RMS) $\frac{1}{t}\sum_{\tau=1}^{t}\{r_{\tau,a_\tau} - m_{\mu,\theta}(b_{\tau,a_\tau})\}^2$ for (8), which is computed on the context and reward pair of the chosen arms. Let $I_i(\tau) = I(a_\tau = i)$ be the binary indicator taking value 1 if $a_\tau = i$ and 0 otherwise. The expectation of the RMS is

$$\begin{aligned}\mathbb{E}\left[\text{RMS}\right] &= \mathbb{E}\left[\frac{1}{t}\sum_{\tau=1}^{t}\sum_{i=1}^{N}\{r_{\tau,i} - m_{\mu,\theta}(b_{\tau,i})\}^2 I_i(\tau)\right] \\ &= \mathbb{E}\left[\frac{1}{t}\sum_{\tau=1}^{t}\sum_{i=1}^{N}\{r_{\tau,i} - m_{\mu,\theta}(b_{\tau,i})\}^2\mathbb{E}[I_i(\tau)|\mathcal{F}_{\tau-1}]\right] \\ &= \mathbb{E}\left[\frac{1}{t}\sum_{\tau=1}^{t}\sum_{i=1}^{N}\{r_{\tau,i} - m_{\mu,\theta}(b_{\tau,i})\}^2\pi_{\hat{\theta}_{\tau-1}}(b_\tau,i)\right],\end{aligned}$$

where $\mathcal{F}_{t-1}$ denotes a filtration at time t, the union of the history $\mathcal{H}_{t-1}$ of observations up to time $t-1$ and the context b_t at time t, i.e., $\mathcal{F}_{t-1} = \mathcal{H}_{t-1}\cup\{b_t\}$ where $\mathcal{H}_{t-1} = \bigcup_{\tau=1}^{t-1}\{b_\tau, a_\tau, r_{\tau,a_\tau}\}$. Due to Azuma-Hoeffding's inequality, the RMS converges in probability to $\mathbb{E}[\text{RMS}]$ for any μ and θ. A gap with (8) is caused by the udpate of $\hat{\theta}_\tau$ at each time point. To resolve this, we propose to minimize the following importance-weighted RMS instead,

$$\hat{U}^t(\mu,\theta) = \frac{1}{t}\sum_{\tau=1}^{t}\{r_{\tau,a_\tau} - m_{\mu,\theta}(b_{\tau,a_\tau})\}^2\frac{\pi_\theta(b_\tau,a_\tau)}{\pi_{\hat{\theta}_{\tau-1}}(b_\tau,a_\tau)} \quad (9)$$

$$= \frac{1}{t}\sum_{\tau=1}^{t}\sum_{i=1}^{N}\{r_{\tau,i} - m_{\mu,\theta}(b_{\tau,i})\}^2\pi_\theta(b_\tau,i)\frac{I_i(\tau)}{\pi_{\hat{\theta}_{\tau-1}}(b_\tau,i)} \quad (10)$$

Since $\mathbb{E}[I_i(\tau)|\mathcal{F}_{\tau-1}] = \pi_{\hat{\theta}_{\tau-1}}(b_\tau,i)$, the expectation of $\hat{U}^t(\mu,\theta)$ is exactly (8), and $\hat{U}^t(\mu,\theta)$ converges in probability to (8) for any μ and θ.

We note here that if we had the guarantee that $\hat{\theta}_t$ converges in probability to θ^*, then the RMS would converge to (8) as well. However, the convergence of $\hat{\theta}_t$ to θ^*

Algorithm 2: Actor-Improper Critic algorithm

1: Set $\lambda > 0$, $C > 1$, $\hat{\theta}_0 = 0_d$.
2: **for** $t = 1, \cdots, T$ **do**
3: Pull arm a_t according to probability $\left\{\pi_{\hat{\theta}_{t-1}}(b_t, i)\right\}_{i=1}^{N}$ and get reward r_{t,a_t}.
4: **Critic update**: $\hat{\mu}_t \leftarrow \underset{\mu:||\mu||_2 \leq C}{\operatorname{argmin}} \hat{U}^t(\mu, \hat{\theta}_{t-1})$ (see (10))
5: **Actor update**: $\hat{\theta}_t \leftarrow \underset{\theta}{\operatorname{argmax}} \hat{J}^t(\hat{\mu}_t, \theta)$ (see (11)).
6: **end for**

is guaranteed only when the compatibility condition holds, which requires itself that the RMS converges to (8).

The empirical version of (7) is

$$\hat{J}^t(\mu, \theta) = \frac{1}{t} \sum_{\tau=1}^{t} \sum_{i=1}^{N} b_{\tau,i}^T \mu \, \pi_\theta(b_\tau, i) - \lambda \theta^T \theta, \qquad (11)$$

and its expectation is exactly (7). In the next section, we prove that the values of μ and θ that minimize $\hat{U}^t(\mu, \theta)$ with respect to μ and maximize $\hat{J}^t(\mu, \theta)$ with respect to θ converge in probability to μ^* and θ^* respectively.

Computation

The proposed algorithm with the new estimating functions is presented in Algorithm 2. At each iteration of the algorithm, we find the value $\hat{\mu}_t$ which minimizes $\hat{U}^t(\mu, \theta)$ with θ replaced with $\hat{\theta}_{t-1}$ from the previous iteration. Then we find the value $\hat{\theta}_t$ which maximizes $\hat{J}^t(\mu, \theta)$ with μ replaced with $\hat{\mu}_t$. The inverse probability $1/\pi_{\hat{\theta}_{\tau-1}}(b_\tau, i)$ can have large value and cause instability of the estimate $\hat{\mu}_t$. To mitigate such instability, we solve $\hat{\mu}_t = \underset{\mu:||\mu||_2 \leq C}{\operatorname{argmin}} \hat{U}^t(\mu, \theta)$ for some positive constant C. We later show that if C is set such that $\mu^* \in \{\mu : ||\mu||_2 \leq C\}$, $\hat{\mu}_t$ and $\hat{\theta}_t$ converge in probability to μ^* and θ^* respectively. Without the constraint, $\hat{\mu}_t$ is just a weighted least-squares estimator with importance weights $\pi_{\hat{\theta}_t}(b_\tau, a_\tau)/\pi_{\hat{\theta}_{\tau-1}}(b_\tau, a_\tau)$'s. We later show that $\hat{\mu}_t$ with the constraint converges to the weighted least-squares estimator as time accumulates.

Regret Bound

The proposed algorithm (Algorithm 2) is robust to the misspecification of the critic model and converges to the optimal action selection policy. We define the regret with respect the optimal action selection policy as follows,

$$R(T) = \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{E}[r_{t,i}|b_{t,i}] \left\{\pi_{\theta^*}(b_t, i) - \pi_{\hat{\theta}_{t-1}}(b_t, i)\right\}.$$

We can show that the proposed algorithm achieves a regret that is upper-bounded by $O(\sqrt{T})$ with high probability. This upper bound is of same order as the regret upper bound of Algorithm 1 which requires a restrictive assumption that the linear model correctly specifies the reward model. We provide the proofs in the Supplementary Material.

Asymptotic Properties and Testing Procedure

Statistical tests on the significance of the true parameter values (μ^* and θ^*) can be conducted if the distribution of the estimates are known. In this section, we derive the asymptotic distribution of $\hat{\mu}_t$ and $\hat{\theta}_t$. We first state some necessary assumptions.

Assumption 1. *The distribution of contexts variables is i.i.d. over time t, i.e.,*

$$b_t = \{b_{t,1}, \cdots, b_{t,N}\} \overset{i.i.d.}{\sim} P_b,$$

where P_b is some distribution over $\mathbb{R}^{N \times d}$. Also, the distribution of rewards $r_t = \{r_{t,1}, \cdots, r_{t,N}\}$ is i.i.d. over time t.

Assumption 2. *Contexts and rewards are bounded. Without loss of generality, $||b_{t,i}||_2 \leq 1$ and $|r_{t,i}| \leq 1$.*

Assumption 3. *The optimal policy θ^* is unique and μ^* is unique, and the joint equation $\left[\left\{\frac{\partial}{\partial \mu} U(\mu, \theta)\right\}^T, \left\{\frac{\partial}{\partial \theta} J(\mu, \theta)\right\}^T\right] = 0_{2d}^T$ has unique solution at $[\mu^T, \theta^T] = [\mu^{*T}, \theta^{*T}]$. Moreover, for a fixed value of μ, $J(\mu, \theta)$ has unique maximum at $\theta = \theta_\mu^*$. Also for a fixed value of θ, $U(\mu, \theta)$ has unique minimum at $\mu = \mu_\theta^*$.*

Assumption 4. *Let $\bar{b}_\theta(t) = \sum_{i=1}^{N} \pi_\theta(b_t, i) b_{t,i}$. The matrix $\mathbb{E}_\theta[(b_{t,a_t} - \bar{b}_\theta(t))(b_{t,a_t} - \bar{b}_\theta(t))^T] = \mathbb{E}\left[\sum_{i=1}^{N} \pi_\theta(b_t, i)(b_{t,i} - \bar{b}_\theta(t))(b_{t,i} - \bar{b}_\theta(t))^T\right]$ is positive definite for θ in a neighborhood of θ^*.*

Assumption 1 is standard in literature (Langford and Zhang 2007; Goldenshluger and Zeevi 2013; Bastani and Bayati 2020) and is reasonable in many practical settings such as clinical trials where arms have a stationary distribution and do not depend on the past. The uniqueness of μ^* follows under mild conditions as it minimizes a convex function (8) and because we can discard all the contextual features which do not differ by arms. The uniqueness of θ^* is a reasonable assumption as well since the penalty $-\lambda\theta^T\theta$ in (2) introduces a degree of convexity. Also due to this penalty $||\theta^*||_2$ is bounded, so the optimal policy itself is a policy that enforces a positive probability (γ) of uniform exploration. Therefore, we have $\mathbb{E}_{\theta^*}[(b_{t,a_t} - \bar{b}_{\theta^*}(t))(b_{t,a_t} - \bar{b}_{\theta^*}(t))^T] \succeq \gamma \mathbb{E}\left[\sum_{i=1}^{N} (b_{t,i} - \bar{b}_{\theta^*}(t))(b_{t,i} - \bar{b}_{\theta^*}(t))^T\right]$. Positive-definiteness of the right-hand side imposes variety in the arm features and is also a standard assumption in the literature. (See Goldenshluger and Zeevi (2013) and Bastani and Bayati (2020).) Hence, Assumption 4 is reasonable as well.

We first show the following lemma which is crucial in deriving the consistency of the estimates.

Lemma 2. *Under Assumptions 2-4, the optimal policy parameter θ^* and μ^* lie in a compact set. Also, the estimated parameters $\hat{\mu}_t$ and $\hat{\theta}_t$ lie in a compact set for all $t \in [T]$.*

Sketch of Proof. We first show the boundedness of μ^*.

Since μ^* is the minimizer of $U(\mu, \theta^*)$, we have

$$\mu^* = \left\{ \mathbb{E}\left[\sum_{i=1}^{N} \pi_{\theta^*}(b_t, i)(b_{t,i} - \bar{b}_{\theta^*}(t))(b_{t,i} - \bar{b}_{\theta^*}(t))^T \right] \right\}^{-1} \times \mathbb{E}\left[\sum_{i=1}^{N} \pi_{\theta^*}(b_t, i)(b_{t,i} - \bar{b}_{\theta^*}(t)) r_{t,i} \right].$$

Due to Assumption 2, we have $||\mu^*||_2 \leq 1/\phi^2$ where

$$\phi^2 = \lambda \left(\mathbb{E}\left[\sum_{i=1}^{N} \pi_{\theta^*}(b_t, i)(b_{t,i} - \bar{b}_{\theta^*}(t))(b_{t,i} - \bar{b}_{\theta^*}(t))^T \right] \right)$$

and $\lambda(\cdot)$ denotes the minimum eigenvalue. Due to Assumption 4 $\phi^2 > 0$, so μ^* lies in a compact set. Now, since θ^* maximizes $J(\mu^*, \theta)$, we have $J(\mu^*, 0_d) \leq J(\mu^*, \theta^*)$. Due to $||\mu^*||_2 \leq 1/\phi^2$, Assumption 2, and Cauchy-Schwarz inequality, we have $J(\mu^*, \theta^*) \leq \frac{1}{\phi^2} - \lambda \theta^{*T}\theta^*$. Also, $J(\mu^*, 0_d) \geq -\frac{1}{\phi^2}$. Therefore, we have $-\frac{1}{\phi^2} \leq \frac{1}{\phi^2} - \lambda\theta^{*T}\theta^*$ which shows that $||\theta^*||_2 \leq \sqrt{\frac{2}{\lambda\phi^2}}$. Due to line 4 of Algorithm 2, $||\hat{\mu}_t||_2 \leq C$ so $\hat{\mu}_t$ clearly lies in a compact set, and analogously, we can show that $||\hat{\theta}_t||_2 \leq \sqrt{\frac{2C}{\lambda\phi^2}}$.

We now prove in Theorem 1 the consistency of the estimates $\hat{\mu}_t$ and $\hat{\theta}_t$.

Theorem 1. Consistency *Let $C^* = 1/\phi^2$. Under assumptions 1-4, if $C^* \leq C$, $(\hat{\mu}_t^T, \hat{\theta}_t^T)$ converges to (μ^{*T}, θ^{*T}) in probability.*

Sketch of Proof. We denote $\Omega = \{(\mu^T, \theta^T) : ||\mu||_2 \leq C, ||\theta||_2 \leq 2\sqrt{2C/(\lambda\phi^2)}\}$. Then Ω forms a compact set and includes (μ^{*T}, θ^{*T}). Since $\sum_{i=1}^{N} b_{\tau,i}^T \mu\, \pi_\theta(b_\tau, i)$ is i.i.d. over time τ for fixed μ and θ, we can apply Glivenko-Cantelli Theorem to $\hat{J}^t(\mu, \theta)$ and prove uniform convergence.

$$\sup_{(\mu^T, \theta^T) \in \Omega} \left| \hat{J}^t(\mu, \theta) - J(\mu, \theta) \right| \xrightarrow{P} 0. \tag{12}$$

Due to the term $I_i(\tau)/\pi_{\hat{\theta}_{\tau-1}}(b_\tau, i)$, $\hat{U}^t(\mu, \theta)$ is not the mean of i.i.d. variables and requires additional steps to prove the uniform convergence. Define

$$\tilde{U}^t(\mu, \theta) = \frac{1}{t} \sum_{\tau=1}^{t} \sum_{i=1}^{N} \{r_{\tau,i} - m_{\mu,\theta}(b_{\tau,i})\}^2 \pi_\theta(b_\tau, i).$$

Using martingale inequalities along with a covering argument on the space Ω, we first show that $|\hat{U}^t(\mu, \theta) - \tilde{U}^t(\mu, \theta)|$ converges uniformly to 0 in probability. Then we apply Glivenko-Cantelli theorem to $\tilde{U}^t(\mu, \theta)$ to finally prove

$$\sup_{(\mu^T, \theta^T) \in \Omega} \left| \hat{U}^t(\mu, \theta) - U(\mu, \theta) \right| \xrightarrow{P} 0, \tag{13}$$

Since $(\hat{\mu}_t^T, \hat{\theta}_t^T)$ lies in Ω (Lemma 2), $U(\mu, \theta)^T$ and $J(\mu, \theta)^T$ are continuous on Ω, and (μ^{*T}, θ^{*T}) is unique (Assumption 3), we can apply Theorem 9.4 in Keener (2010) to show $(\hat{\mu}_t^T, \hat{\theta}_t^T) \xrightarrow{P} (\mu^{*T}, \theta^{*T})$. Detailed proofs are presented in the Supplementary Material.

Lemma 3. *Suppose $C^* \leq C$. As $t \to \infty$, $\hat{\mu}_t$ converges in probability to the solution of $\hat{U}_\mu^t(\mu, \hat{\theta}_{t-1}) = 0$, where $\hat{U}_\mu^t(\mu, \theta) = \frac{\partial}{\partial \mu}\hat{U}^t(\mu, \theta)$ and $\hat{J}_\theta^t(\mu, \theta) = \frac{\partial}{\partial \theta}\hat{J}^t(\mu, \theta)$.*

Sketch of Proof. We just need to show that the solution $\tilde{\mu}_t$ of $\hat{U}_\mu^t(\mu, \hat{\theta}_{t-1}) = 0$ satisfies $P(\tilde{\mu}_t \leq C) \underset{t\to\infty}{\longrightarrow} 1$, i.e., $P(\hat{\mu}_t = \tilde{\mu}_t) \to 1$. Note that the solution of $\hat{U}_\mu^t(\mu, \hat{\theta}_{t-1}) = 0$ is a weighted least-squares estimator with weights $w_\tau = \pi_{\hat{\theta}_{t-1}}(b_\tau, a_\tau)/\pi_{\hat{\theta}_{\tau-1}}(b_\tau, a_\tau)$, covariates $b_{\tau,a_\tau} - \bar{b}_{\hat{\theta}_{t-1}}(\tau)$, and outcomes r_{τ,a_τ}. The lemma holds due to Assumption 2, 4, and the consistency of $\hat{\theta}_{t-1}$. Detailed proof can be found in the Supplementary Material.

Theorem 2. Asymptotic Normality *Under assumptions 1-4, if $C^* \leq C$, $\sqrt{t}\left(\begin{pmatrix} \hat{\mu}_t \\ \hat{\theta}_t \end{pmatrix} - \begin{pmatrix} \mu^* \\ \theta^* \end{pmatrix} \right)$ converges in distribution to a multivariate normal distribution with mean 0 and variance $\Psi = \Lambda^{-1} V^* \Lambda^{-1}$ where*

$$\Lambda = \begin{bmatrix} U_{\mu\mu}^* & U_{\mu\theta}^* \\ J_{\theta\mu}^* & J_{\theta\theta}^* \end{bmatrix},$$

$$V^* = \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbb{E}\left[\begin{bmatrix} u_\mu^{\tau *} u_\mu^{\tau *T} & u_\mu^{\tau *} j_\theta^{\tau *T} \\ j_\theta^{\tau *} u_\mu^{\tau *T} & j_\theta^{\tau *} j_\theta^{\tau *T} \end{bmatrix} \middle| \mathcal{F}_{\tau-1} \right],$$

$$u_\mu^\tau(\mu, \theta) = -\sum_{i=1}^{N} 2\{r_{\tau,i} - m_{\mu,\theta}(b_{\tau,i})\} \dot{m}_{\mu,\theta}(b_{\tau,i}) \times \frac{I_i(\tau)}{\pi_{\hat{\theta}_{\tau-1}}(b_\tau, i)} \pi_\theta(b_\tau, i),$$

$$j_\theta^\tau(\mu, \theta) = \sum_{i=1}^{N} b_{\tau,i}^T \mu \dot{\pi}_\theta(b_\tau, i) - 2\lambda\theta,$$

$U_{\mu\mu}$ and $U_{\mu\theta}$ are second order partial derivatives of U with respect to μ twice and with respect to μ and θ respectively, the $J_{\theta\mu}$ and $J_{\theta\theta}$ are defined analogously, and the $U_{\mu\mu}^*, U_{\mu\theta}^*, J_{\theta\mu}^*, J_{\theta\theta}^*, u_\mu^{\tau *}, j_\theta^{\tau *}$ are the values of $U_{\mu\mu}, U_{\mu\theta}, J_{\theta\mu}, J_{\theta\theta}, u_\mu^\tau, j_\theta^\tau$ evaluated at the true value (μ^{*T}, θ^{*T}). The asymptotic variance Ψ can be estimated by replacing the expectation operation $\mathbb{E}(\cdot)$ with the empirical mean and plugging-in the estimates $(\hat{\mu}_t^T, \hat{\theta}_t^T)$. Due to Assumption 1 and consistency of $(\hat{\mu}_t^T, \hat{\theta}_t^T)$, such plug-in type estimator is consistent for the asymptotic variance.

Sketch of Proof. Due to Lemma 3 and linearization method, for sufficiently large t,

$$\begin{bmatrix} 0_d \\ 0_d \end{bmatrix} = \begin{bmatrix} \hat{U}_\mu^t(\hat{\mu}_t, \hat{\theta}_t) \\ \hat{J}_\theta^t(\hat{\mu}_t, \hat{\theta}_t) \end{bmatrix} = \begin{bmatrix} \hat{U}_\mu^t(\mu^*, \theta^*) \\ \hat{J}_\theta^t(\mu^*, \theta^*) \end{bmatrix} + \begin{bmatrix} \hat{U}_{\mu\mu}^t(\tilde{\mu}, \tilde{\theta}) & \hat{U}_{\mu\theta}^t(\tilde{\mu}, \tilde{\theta}) \\ \hat{J}_{\theta\mu}^t(\breve{\mu}, \breve{\theta}) & \hat{J}_{\theta\theta}^t(\breve{\mu}, \breve{\theta}) \end{bmatrix} \begin{pmatrix} \hat{\mu}_t - \mu^* \\ \hat{\theta}_t - \theta^* \end{pmatrix}$$

where $\tilde{\mu} = \alpha\hat{\mu}_t + (1-\alpha)\mu^*, \tilde{\theta} = \alpha\hat{\theta}_t + (1-\alpha)\theta^*, \breve{\mu} = \beta\hat{\mu}_t + (1-\beta)\mu^*$ and $\breve{\theta} = \beta\hat{\theta}_t + (1-\beta)\theta^*$ for some $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$. Due to the consistency of $(\hat{\mu}^T, \hat{\theta}^T)$ and the

Param.	ε-greedy		Param.	AC	Proposed
	$i=1$	$i=2$			
$\mu^i_{(i-1)d+1}$	0.91	0.93	θ_1	0.77	0.99
$\mu^i_{(i-1)d+2}$	0.93	0.94	θ_2	0.64	0.99
$\mu^i_{(i-1)d+3}$	0.96	0.90	θ_3	0.74	0.99
$\mu^i_{(i-1)d+4}$	0.58	0.59	θ_4	0.07	0.09

Table 1: Rejection rates of H_0 for each parameter (Param.) by ε-greedy (Chen, Lu, and Song 2020), Actor-Critic (Lei, Tewari, and Murphy 2017), and Proposed algorithm

Law of Large Numbers,

$$\sqrt{t}\begin{pmatrix}\hat{\mu}_t-\mu^*\\ \hat{\theta}_t-\theta^*\end{pmatrix}=-\left\{\begin{bmatrix}U^*_{\mu\mu} & U^*_{\mu\theta}\\ J^*_{\theta\mu} & J^*_{\theta\theta}\end{bmatrix}+o_P(1)\right\}^{-1}\times\sqrt{t}\begin{bmatrix}\hat{U}^t_\mu(\mu^*,\theta^*)\\ \hat{J}^t_\theta(\mu^*,\theta^*)\end{bmatrix}$$

Since the $\hat{J}^t_\theta(\mu^*,\theta^*)$ is the empirical mean of i.i.d. variables with mean 0, we can apply the Central Limit Theorem (CLT) to derive the asymptotic distribution. On the other hand, $\hat{U}^t_\mu(\mu^*,\theta^*)$ is the empirical mean of $u^\tau_\mu(\mu^*,\theta^*)$ which are not i.i.d. due to the term $I_i(\tau)/\pi_{\hat{\theta}_{\tau-1}}(b_\tau,i)$. Instead, the $u^\tau_\mu(\mu^*,\theta^*)$'s form a martingale difference sequence. Hence, we can apply martingale CLT to $\sqrt{t}\left[\hat{U}^t_\mu(\mu^*,\theta^*)^T\hat{J}^t_\theta(\mu^*,\theta^*)^T\right]$ in whole and show that this converges to a normal distribution with mean 0 and variance V^*.

Based on Theorem 2, a Z-test can be conducted for a j-th variable ($j=1,\cdots,d$) using the test statistic $Z=\hat{\theta}_{t,j}/\sqrt{\Psi_{d+j,d+j}/t}$. We reject the null hypothesis $H_0:\theta_j=0$ with significance level α when $2(1-\Phi(|Z|))<\alpha$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

Experiments

We conduct experiments to evaluate the performance of the Proposed algorithm under a misspecified reward model. We set $N=2$ and $d=4$. We generate the context vectors $b_{t,i}$ from a multivariate normal distribution $\mathcal{N}(0_d, I^{d\times d})$ and truncate them to have L_2-norm 1. We generate the reward from a model nonlinear in $b_{t,i}$, $r_{t,i}=b^T_{t,i}\mu-\max(b^T_{t,1}\mu,b^T_{t,2}\mu)+\eta_{t,i}$ where $\mu=(-0.577,0.577,0.577,0)^T$ and $\eta_{t,i}$ is generated from $\mathcal{N}(0,0.01^2)$ independently over arms and time. To test the validity of the proposed testing procedure, we set μ_4 to 0 so that the corresponding variable does not affect the reward and hence, will not be useful in the policy.

We implement the Proposed algorithm (Actor-Improper Critic) along with the original Actor-Critic algorithm (Lei, Tewari, and Murphy 2017) and the ε-greedy algorithm using weighted LSE (Chen, Lu, and Song 2020). When implementing the Proposed algorithm, we drop the restriction on $||\hat{\mu}_t||_2$ and compute the weighted least-squares estimator for simplicity. Since the objective function $J(\mu,\theta)$ is not convex with respect to θ, we find the maximizer through a grid search followed by pattern search based on the Nelder-Mead method (Nelder and Mead 1965) as suggested in Lei, Tewari, and Murphy (2017). When implementing the ε-greedy algorithm, we stack the context vectors into one vector $b^T_t=[b^T_{t,1},\cdots,b^T_{t,N}]$ and set the working model for the reward of the i-th arm as $f_i(b_t)=b^T_t\mu^i$, where μ^i is a parameter of dimension Nd. We set the exploration parameter λ in the AC and Proposed algorithms to 0.001. In the ε-greedy algorithm, we use the value $\varepsilon=0.01$ which corresponds to the exploration probability guaranteed by $\lambda=0.001$ in the Proposed algorithm. We run the bandit algorithms until time horizon $T=50$ with 100 repetitions.

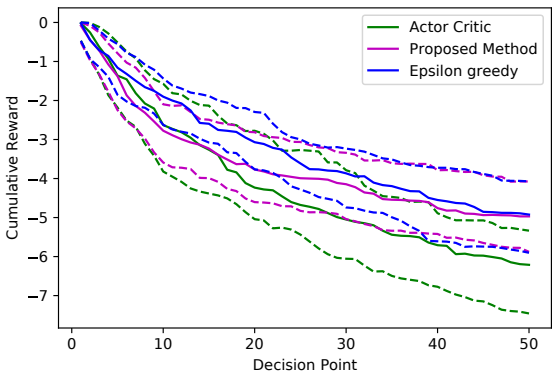

Figure 1: Median (solid lines) and first and third quartiles (dashed lines) of the cumulative rewards.

For each algorithm, we count the number of times the null hypotheses $H_0:\theta_j=0$ (for AC and Proposed algorithms) or $H_0:\mu^i_{(i-1)d+j}=0$ (for ε-greedy) are rejected at time T according to a Z-test with significance level $\alpha=0.05$. We report the rejection rates of H_0 in Table 1. We observe that the AC algorithm (Lei, Tewari, and Murphy 2017) fails to reject the null hypotheses for the non-zero parameters θ_1,θ_2, and θ_3 for at least 23% of the experiments. On the other hand, the Proposed algorithm rejects the null hypotheses 99 times out of 100 times. As for the fourth parameter which has true value 0, both algorithms reject the null hypothesis with small probability. We note that the significance level α lies in the 95% confidence interval $[0.047, 0.133]$ computed from the formula $\left[\hat{p}-1.96\sqrt{\alpha(1-\alpha)/n},\hat{p}+1.96\sqrt{\alpha(1-\alpha)/n}\right]$, where $\hat{p}=0.09$ is the rejection rate of H_0 by the Proposed algorithm and $n=100$ is the number of experiments. On the other hand, the ε-greedy algorithm rejects the null hypothesis for the fourth parameter with more than 50% frequency. We also note that the power of the tests for the ε-greedy algorithm with weighted LSE is lower than the Proposed algorithm. One possible reason for the low

	AC	Proposed	ε-greedy
Mean	5296.9	5557.7	5577.3
St.d.	139.3	152.7	274.9

Table 2: Mean and standard deviations (St.d.) of cumulative rewards at $T = 1000$ by ε-greedy (Chen, Lu, and Song 2020), Actor-Critic (Lei, Tewari, and Murphy 2017), and Proposed algorithm for the Recovery Record Dataset

performance in testing is due to the variance induced by inverse probability weighting. The Proposed algorithm also involves computation of the inverse of probabilities, but it is used only in the ratio of probabilities $\pi_{\hat{\theta}_t}(b_\tau, i)/\pi_{\hat{\theta}_{\tau-1}}(b_\tau, i)$ which converges to 1 as $\hat{\theta}_{\tau-1}$ converges. As for the cumulative rewards, we observe that the Proposed and the ε-greedy algorithm show comparable performance.

Data Application

The Recovery Record Dataset contained patients' adherence behaviors to their therapy for eating disorders (daily meal monitoring) and interactions with their linked clinicians on the app. Clinician communication is often viewed as a critical means to encourage adherence to monitoring, yet there is little guidance of when and how clinicians should communicate outside of office visits, and thus is done on an ad-hoc and individual basis. The rewards (i.e. whether the patient adhered to daily monitoring) were observed for the actions chosen by the ad-hoc policies of clinicians (i.e. send a message or not). A contextual bandit algorithm can allow clinicians to tailor their communications (arms) with patients who have preferences (contexts) to maximize adherence to therapy (rewards).

We applied the offline policy evaluation method of Li et al. (2011) to unbiasedly estimate the cumulative reward that we would obtain under the AC, Proposed, and ε-greedy algorithms. We repeated the evaluations on 30 bootstrap samples. Table 2 shows the mean and standard deviations of the cumulative rewards at time $T = 1000$. We remark that although the Proposed and ε-greedy algorithms have comparable cumulative rewards, ε-greedy suffers higher variance. More details on the implementation and testing results can be found in the Supplementary Material. We show in the Supplementary material the rejection rates of H_0 for ε-greedy algorithms are closer to 0.5 as compared to the Proposed algorithm, which implies that ε-greedy may have lower power due to high variance.

Conclusion

The problem of testing the utility of variables collected by wearables or sensors represents a practical need in mHealth and is an inferential problem highlighted by Tewari and Murphy (2017). In this paper, we considered testing the utility of variables in the actor-critic bandit, but considered inference when the models used in the algorithms are not correctly specified. This work demonstrates that a robust test can be constructed for the actor parameter. Such work also illustrates that inferential procedures associated with the actor-critic bandit inherit problems such as model misspecification by virtue of the assumptions made in model-based testing; however, existing tools in the literature do not apply due to the unique structure of the objective function. This paper adds to the literature by developing a new statistical procedure to test the actor parameters in the actor-critic bandit even when the reward model is misspecified by the critic.

Strengths of this study include its contribution to model-based estimation in the context of contextual bandit algorithms, and the key property that these results are robust even when the assumptions underlying the critic fail to be correct.

The ability to test variables may offer guidance in understanding what types of data (e.g. location or sensitive information) are useful. Beyond the computational and performance related impact of this work, such knowledge could have societal impact in that it may discourage unnecessary data collection, thereby mitigating potential risks and threats to privacy. On the other hand, this method may also provide supporting evidence that certain types of data, perhaps either sensitive or costly variables, are in fact useful. In such cases, the ethical tension between acquiring sensitive information or choosing not to acquire it at the cost of sub-optimal decision-making deserves disclosure and careful discussion with stakeholders involved in and affected by the algorithm. This work provides a means to gauge the significance and assumptions made about the utility of data that would otherwise go untested, as it commonly does at present.

While the focus of this paper is on testing, future work that focuses on implementation aspects of this procedure would be beneficial, such as addressing the question of when these tests should be performed in practice. The proposed method suggests that one feasible method to conduct testing after a large number of trials, but alternatives such as sequential or repeated hypothesis testing merit further attention.

Acknowledgments

Gi-Soo Kim is supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2020-0-01336, Artificial Intelligence Graduate School Program(UNIST)) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT, No.2021R1G1A100980111). Gi-Soo Kim is also partly funded by the 0000 Project Fund (Project Number 1.210079.01) of UNIST, South Korea.

Jane Paik Kim and Hyun-Joon Yang are supported by the National Institutes of Health Grant R01TR003505.

References

Baird III, L. C. 1993. Advantage updating. Technical Report ADA280862, Defense Tech. Inform. Center.

Bastani, H.; and Bayati, M. 2020. Online decision making with high-dimensional covariates. *Operations Research*, 68(1): 276–294.

Chen, H.; Lu, W.; and Song, R. 2020. Statistical Inference for Online Decision-Making: In a Contextual Bandit Setting. *Journal of the American Statistical Association*, 116(533): 240–255.

Ghosh, A.; Chowdhury, S. R.; and Gopalan, A. 2017. Misspecified linear bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Goldenshluger, A.; and Zeevi, A. 2013. A linear response bandit problem. *Stochastic Systems*, 3(1): 230–261.

Hao, B.; Abbasi Yadkori, Y.; Wen, Z.; and Cheng, G. 2019. Bootstrapping Upper Confidence Bound. In *Advances in Neural Information Processing Systems*, volume 32, 12123–12133.

Keener, R. 2010. *Theoretical statistics: Topics for a core course*. Springer Science & Business Media.

Kim, J. P. 2013. A Note on Using Regression Models to Analyze Randomized Trials: Asymptotically Valid Hypothesis Tests Despite Incorrectly Specified Models. *Biometrics*, 69(1): 282–289.

Langford, J.; and Zhang, T. 2007. Epoch-Greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems*, volume 20, 1–8.

Lei, H.; Tewari, A.; and Murphy, S. A. 2017. An actor-critic contextual bandit algorithm for personalized mobile health interventions. arXiv:1706.09090.

Li, L.; Chu, W.; Langford, J.; and Wang, X. 2011. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 297–306.

Nelder, J. A.; and Mead, R. 1965. A simplex method for function minimization. *The computer journal*, 7(4): 308–313.

Rosenblum, M.; and Van Der Laan, M. J. 2009. Using regression models to analyze randomized trials: Asymptotically valid hypothesis tests despite incorrectly specified models. *Biometrics*, 65(3): 937–945.

Sutton, R. S.; McAllester, D. A.; Singh, S. P.; and Mansour, Y. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, 1057–1063.

Tang, Q.; Xie, H.; Xia, Y.; Lee, J.; and Zhu, Q. 2021. Robust Contextual Bandits via Bootstrapping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 12182–12189.

Tewari, A.; and Murphy, S. A. 2017. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, 495–517. Springer, Cham.

Zhang, K. W.; Janson, L.; and Murphy, S. A. 2021. Statistical Inference with M-Estimators on Bandit Data. arXiv:2104.14074.

Zhu, F.; Guo, J.; Li, R.; and Huang, J. 2018. Robust actor-critic contextual bandit for mobile health (mhealth) interventions. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 492–501.