An Actor-Critic Contextual Bandit Algorithm for Personalized Mobile Health Interventions

Huitian Lei

Department of Statistics, University of Michigan

Ambuj Tewari

Department of Statistics, University of Michigan

Susan A. Murphy *

Department of Statistics,

Department of Psychiatry,

Institute of Social Research, University of Michigan

June 29, 2017

Abstract

Increasing technological sophistication and widespread use of smartphones and wearable devices provide opportunities for innovative and highly personalized health interventions. A Just-In-Time Adaptive Intervention (JITAI) uses real-time data collection and communication capabilities of modern mobile devices to deliver interventions in real-time that are adapted to the in-the-moment needs of the user. The lack of methodological guidance in constructing data-based JITAIs remains a hurdle in advancing JITAI research despite the increasing popularity of JITAIs among clinical scientists. In this article, we make a first attempt to bridge this methodological gap by formulating the task of tailoring interventions in real-time as a contextual bandit problem. Interpretability requirements in the domain of mobile health lead us to formulate the problem differently from existing formulations intended for web applications such as ad or news article placement. Under the assumption of linear reward function, we choose the reward function (the "critic") parameterization separately from a lower dimensional parameterization of stochastic policies (the "actor"). We provide an online actor-critic algorithm that guides the construction and refinement of a JITAI. Asymptotic properties of the actor-critic algorithm are developed and backed up by numerical experiments. Additional numerical experiments are conducted to test the robustness of the algorithm when idealized assumptions used in the analysis of contextual bandit algorithm are breached.

1 Introduction

Equipped with sophisticated sensing, communication and computation capabilities, smartphones and mobile devices are being increasingly used to deliver Just-In-Time Adaptive Interventions (JITAIs). JITAIs are mobile health interventions where treatment is delivered in real time to individuals as they go about their daily lives. A key ingredient of a JITAI is a *policy*, that is, a decision rule that inputs sensor and self-report information at any given decision point and output a decision. The decision can be whether or not to provide treatment or the type of treatment to be provided. The use of decision rules to adapt the type and timing of treatment delivery to the individual makes JITAIs particularly promising in facilitating *long-term* health behavior change, a pressing but notoriously hard problem (Nahum-Shani et al. (2014)). Indeed JITAIs have received increasing popularity and have been used to support health behavior change in a variety of domains including physical activity (King et al. (2013); Consolvo et al. (2008)), eating disorders (Bauer et al. (2010)), drug abuse (Scott and Dennis (2009)), alcohol use (Witkiewitz et al. (2014); Suffoletto et al. (2012); Gustafson et al. (2011)), smoking cessation (Riley et al. (2011)), obesity and weight management (Patrick et al. (2009)), and other chronic disorders.

Despite the growing popularity of JITAIs, there is a lack of guidance concerning how to best learn a high-quality evidence-based JITAIs in an "online" setting. That is, learning occurs in a sequential manner as a given user experiences the treatments and sensor/self-report data, including health outcomes of interest, are collected. Ideally, the policy we learn for a given user should take into account the specific way he or she responds to the delivered treatments and is thus *personalized* to the user. However, most of the JITAIs used in existing clinical trials are specified a priori and are based primarily on domain expertise. The main contribution of this article to take a first step towards bridging the gap between the enthusiasm for JITAIs in the mobile health field and the current lack of statistical methodology to guide the online construction of a personalized policy for a user. We model the learning of a user-specific optimal policy as a contextual bandit problem (Woodroofe (1979); Langford and Zhang (2008); Li et al. (2010)). A contextual bandit problem, also called a bandit problem with side-information, is a sequential decision making

problem where a learning algorithm, (i) chooses an action (e.g., treatment) at each time point based on the *context* or side information, and (ii) receives an reward that reflects the quality of the action under the current context. In mobile health settings, the context can include summaries of the sensor and self-report data available at each time point. The goal of the algorithm is to learn the optimal policy, that is, the policy that maximizes a regularized average reward for a user. We propose an online "actor-critic" algorithm for learning the optimal policy. Compared to offline learning, in online learning the contexts and rewards arrive in a sequential fashion and the estimate of the optimal policy is updated as data accumulates. The updated policy is used to choose the treatment action at the subsequent time point. In our actor-critic algorithm, the critic estimates parameters in a model for the conditional mean of the reward given context and action. The actor then updates the estimated optimal policy based on the estimated reward model. Under idealized assumptions, we derive asymptotic theory for the consistency and asymptotic normality of the estimated optimal policy.

Our work is motivated by our collaboration on HeartSteps (Klasnja et al. (2015); Dempsey et al. (2015)). In the HeartSteps project, the second and third of three studies will involve the use of a online learning algorithm for constructing personalized policies; the algorithm presented here represents our first step in developing the learning algorithm. The goal of the HeartStepsproject is to reduce sedentary behavior and increase physical activity in individuals who have experienced a cardiac event and been in cardiac rehab. The current version of HeartSteps involves data collection both via a smartphone as well as wristband sensor. A variety of sensor and self-report data is available at each time point, including step count, GPS location, weather, time and user calendar busyness. The current version of HeartSteps can deliver a treatment (an activity suggestion) at any of 5 time points per day via an audible ping and a notification on the smartphone lock screen.

This article is organized as follows. In Section 2, we formulate online learning of a policy for a given user as a contextual bandit problem and define what we mean by an optimal policy. Due to the concern that deterministic policies may habituate users to treatments, thereby causing them to ignore treatment, our definition of optimality is different from the ones found

in most existing contextual bandit papers. In Section 3, we present an actor-critic contextual bandit algorithm for learning the optimal policy. In Section 4, we derive asymptotic theory on the consistency and asymptotic normality of the estimated optimal policy. In Section 5, we present a comprehensive simulation study to investigate the performance of the actor-critic algorithm under various simulation settings including settings which violate the usual assumptions underpinning contextual bandit algorithms.

2   Learning JITAIs via a Contextual Bandit Algorithm

We formulate the online learning of optimal policy for a given user as a *stochastic* contextual bandit problem. A contextual bandit problem is specified by a quadruple $(\mathcal{S}, d, \mathcal{A}, r)$, where $\mathcal{S}$ is the context space, d is a probability distribution on the context space, $\mathcal{A}$ is the action space and r is the reward space. At a decision point t, the online learning algorithm collects the context $S_t \in \mathcal{S}$, take an action $A_t \in \mathcal{A}$ after which a reward $R_t \in r$ is revealed before the next decision point. The online learning algorithm does not have access to the rewards that would occur had other actions been taken. Prior to decision point $t + 1$, the algorithm has access to the sequence of tuples $\{(S_\tau, A_\tau, R_\tau)\}_{\tau=1}^{t}$. We make the following assumption.

Assumption 1. *(i.i.d. contexts) Action A_t has a in-the-moment effect on the reward R_t with expected reward function:*

$$\mathbb{E}\left(R_t | S_t = s, A_t = a\right) = r(s, a),$$

but A_t does not affect the distribution of S_τ for $\tau \geq t + 1$. We further assume that contexts S_t are i.i.d. with probability density function $d(s)$.

This assumption matches the conceptual design of many JITAIs well. In fact, intervention options in a JITAI are sometimes referred to as "Ecological Momentary Interventions" (EMIs) or "micro-interventions". Such a terminology emphasizes that the effects of many of the treatments in this domain are expected to be short-lived in nature.

A *(stochastic) policy* is a mapping from the context space to (a probability distribution over) the action space. In JITAIs, policies are used to specify (the probability of) an action given a context. In this article, we focus on a binary action space $\mathcal{A} = \{0, 1\}$ and

a class of parametrized stochastic policies in which $P(A = 1|S = s)$ is parameterized as $\pi_\theta(s, 1) = \frac{e^{g(s)^T\theta}}{1+e^{g(s)^T\theta}}$. Here $g(s)$ is a p-dimensional vector that contains candidate variables that may be useful for decision making. Using the parametrized policy, the way each variable in $g(s)$ influences the choice of action is reflected by the sign and magnitude of the corresponding component in θ. Confidence intervals for and hypothesis testing on the optimal θ can answer scientific questions about the usefulness of a particular contextual variable for decision making. For example, suppose the scientist includes a GPS location based variable as a candidate variable in the policy, yet the confidence interval for the θ coefficient of this variable turns out to contain 0. Then we might omit the sensing of this variable in future because continuously sensing GPS location on smartphones drains the battery. Similarly, self-reported measures on user's emotional states induce user burden. Therefore, if the confidence interval for the θ coefficients of these variables contains 0 we may reduce user burden by omitting their collection.

2.1   The Regularized Average Reward

Empirical evidence and some behavioral science theories indicate that deterministic policies can lead to habituation and that treatment *variety* can increase user engagement and retard habituation (Raynor and Epstein (2001); Epstein et al. (2009, 2011); Wilson et al. (2005)).To maintain treatment variety, we consider stochastic policies. However it turns out that standard definitions of optimality often lead to deterministic policies. For example, a natural and intuitive definition of an optimal policy is a policy that maximizes the average reward:

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

where $d(s)$ is the probability density function of context. The following lemma shows that, in a simple setting where the context space is one-dimensional and finite, there always exists a deterministic optimal policy. The proof of this lemma is provided in the supplementary material section A.

Lemma 1. *Suppose that the context space is discrete and finite, $\mathcal{S} = \{s_1, s_2, ..., s_K\}$. Among*

6

the policies parameterized as $\pi_\theta(s, 1) = \frac{e^{\theta_0 + \theta_1 s}}{1 + e^{\theta_0 + \theta_1 s}}$, there exists a policy that maximizes $V^*(\theta)$ for which $P(\pi_\theta(S, 1) = 0 \text{ or } 1) = 1$.

One way to ensure treatment variety is to introduce a *chance constraint* (also called a "probabilistic constraint"; see, e.g., Prékopa (1995)) that ensures, with high probability over the context distribution, that the probability of taking each treatment action under any policy we consider is bounded sufficiently away from 0. For binary actions the constraint has the form:

$$P(p_0 \leq \pi_\theta(S, 1) \leq 1 - p_0) \geq 1 - \alpha \tag{1}$$

where $0 < p_0 < 0.5$, $0 < \alpha < 1$ are constants controlling the amount of stochasticity. The stochasticity constraint requires that, for at least $(1 - \alpha)100\%$ of the contexts, there is at least p_0 probability to take either of the two available actions.

Maximizing the average reward $V^*(\theta)$ subject to the stochasticity constraint (1) is a chance constrained optimization problem, an active research area in recent years (Nemirovski and Shapiro (2006); Campi and Garatti (2011)). Solving this chance constraint problem, however, involves a major difficulty: constraint (1) is, in general, a non-convex constraint on θ. Moreover, the left hand side of the chance constraint is an expectation of a non-smooth indicator function. Both the non-convexity and the non-smoothness make the optimization problem computationally intractable. We circumvent this difficulty by relaxing constraint (1) to a convex alternative:

$$\theta^T \mathbb{E}[g(S)g(S)^T]\theta = \theta^T [\int_{s \in \mathcal{S}} g(s)g(s)^T d(s)ds]\theta \leq \left(\log(\frac{p_0}{1 - p_0})\right)^2 \alpha, \tag{2}$$

which is obtained by bounding the probability in (1) using Markov's inequality and some algebra. Since the quadratic constraint is derived using an upper bound on the original probability, it is more stringent than the chance constraint and always guarantees *at least* the desired amount of treatment variety.

Instead of solving the quadratic optimization problem that maximizes the average reward $V^*(\theta)$ subject to the quadratic constraint (2), we choose to maximize the corresponding Lagrangian function. Incorporating inequality constraints by using Lagrangian multipliers has

been widely used in reinforcement learning literature to solve constrained Markov decision problem (Borkar (2005); Bhatnagar and Lakshmanan (2012)). Given a Lagrangian multiplier λ, the following Lagrangian function:

$$J_\lambda^*(\theta) = \int_{s \in \mathcal{S}} d(s) \sum_{a \in \mathcal{A}} r(s, a) \pi_\theta(s, a) ds - \lambda \theta^T \mathbb{E}[g(S)g(S)^T] \theta \tag{3}$$

is referred to as the *regularized average reward* in this article. For a fixed value of λ, we define the *optimal policy* to be the policy that maximizes the regularized average reward, namely $\theta_\lambda^* = \operatorname{argmax} J_\lambda^*(\theta)$. Under mild regularity conditions, we show that there is a one-to-one correspondence between quadratic constrained optimization of the average reward (using the constraint (2)) and the unconstrained optimization of the regularized average reward (3). Details can be found in the supplementary material section B. There are two computational advantages of maximizing the regularized average reward as opposed to solving a constrained optimization. First, optimizing the regularized average reward function results in a unique solution even when there is no treatment effect. When the expected reward does not depend on the treatment action, i.e., $\mathbb{E}(R|S = s, A = a) = \mathbb{E}(R|S = s)$, all policies in the feasible set given by the constraint have the same average reward. The regularized average reward function, in contrast, has a unique maximizer at $\theta = \mathbf{0}_{p \times 1}$, a purely random policy that assigns 50% probability to both actions. Therefore, maximizing the regularized average reward gives rise to a 0 estimand when there is no treatment effect. Second, even when the uniqueness of optimal policy is not an issue, maximization of $J_\lambda^*(\theta)$ has computational advantages over maximization of $V^*(\theta)$ under the constraint (2) because the subtraction of the quadratic term $\lambda \theta^T \mathbb{E}[g(S)g(S)^T] \theta$ introduces a degree of concavity to the surface of $J_\lambda^*(\theta)$, thus stabilizing the optimization.

3   The Online Actor-Critic Algorithm

In this section, we propose an online actor-critic algorithm for learning the policy parameter θ_λ^*. We consider a fixed λ in this section and will drop the subscript in θ_λ^* from now on.

Recall that right before decision point $t + 1$, the observed "training data" consists of a stream of triples $\{(S_\tau, A_\tau, R_\tau)\}_{\tau=1}^t$. The θ coefficients in the optimal policy can be estimated

by maximizing an empirical version of the aforementioned regularized average reward:

$$J_\lambda(\theta) = \frac{1}{t} \sum_{\tau=1}^{t} \sum_{a} r(S_\tau, a) \pi_\theta(S_\tau, a) - \lambda \theta^T \left(\frac{1}{t} \sum_{\tau=1}^{t} g(S_\tau) g(S_\tau)^T \right) \theta. \tag{4}$$

The proposed actor-critic algorithm has two parts: the critic estimates the expected reward, r; the estimated expected reward is then plugged into (4), which is then maximized to estimate the parameter, θ, in the optimal policy. The estimated optimal policy is used to select an action at the next decision point.

We develop a critic algorithm based on a linear assumption for the expected reward function.

Assumption 2. *(Linear model assumption) Given context $S_t = s$ and action $A_t = a$, the reward is generated according to the model $R_t = f(s,a)^T \mu^* + \epsilon_t$, where $f(s,a)$ is a k-dimensional reward feature. The error terms ϵ_t are i.i.d. with mean 0 and variance σ^2. In particular, we have $r(s,a) = f(s,a)^T \mu^*$.*

See Section 5 for simulation results concerning the robustness of the developed method to breakdown of this linearity assumption as well as Section 6 for discussion. We further assume that values of the reward, the reward parameter and the reward feature are bounded as follows.

Assumption 3. *(Bounded rewards and features) There exist constants which provide an a.s. upper bound on the absolute value $|R|$ of the reward as well as the norms $|f(S,A)|_2, |\mu^*|_2$ of the reward feature and reward parameter vector. Furthermore, we assume that we know the constant K for which $P[|R| \leq K] = 1$. Without further loss of generality we assume all of these constants are 1 (including $K = 1$).*

Bounded rewards along with bounded features are standard assumption in the bandit literature (see for instance Agrawal and Goyal (2013)). In practice, a known bound on the reward is usually available. For example, consider HeartSteps in which the reward is the number of steps over 30 minutes following time t; there is a generally accepted upper limit on the number of steps a human can take in 30 minutes.

A natural estimator of the reward parameter μ^* is the L_2 penalized least squares estimator:

$$\hat{\mu}_t = \left(\zeta I + \sum_{\tau=1}^{t} f(S_\tau, A_\tau) f(S_\tau, A_\tau)^T \right)^{-1} \sum_{\tau=1}^{t} f(S_\tau, A_\tau) R_\tau \tag{5}$$

where the ζ is the weight on the L_2 penalty. This penalty ensures invertibility of the first term on the right hand side when t is small (note that k, the dimension of the reward feature f does not vary with time t so the penalty term is solely to ensure invertibility). Note that even though we have assumed that the contexts S_τ are i.i.d., this does not meant that the feature vectors, $f(S_\tau, A_\tau)$ are i.i.d. Indeed recall that the actions A_τ are drawn according to the estimated optimal policy at decision point $\tau - 1$, which depends on the entire history at or before decision point $\tau - 1$. This dependency presents challenges in analyzing the actor-critic algorithm; see Section 4.

An additional challenge in analyzing the actor-critic algorithm is getting around an inherent circular dependence: the boundedness of the actor estimates depends on the boundedness of the estimated reward function, or equivalently the critic. The critic's estimates, in turn, depends on the actions selected by the actor. To deal with this challenge we use the known bound from assumption 3, to construct a bounded estimator of $r(s, a)$. The penalized least squares estimator results in the estimator $\hat{r}_t(S, A) = f(S, A)^T \hat{\mu}_t$; this estimator may not be bounded a.s. even though by assumption 3, $|r(S, A)|$ is bounded by 1 a.s. We enforce boundedness on the estimator of $r(s, a)$; in particular we replace $\hat{r}_t(s, a)$ by

$$\hat{r}_t(s, a) = \begin{cases} -2 & \text{if } f(s, a)^T \hat{\mu}_t < -2 \\ f(s, a)^T \hat{\mu}_t & \text{if } |f(s, a)^T \hat{\mu}_t| \le 2 \\ 2 & \text{if } f(s, a)^T \hat{\mu}_t > 2 \end{cases} \tag{6}$$

The estimated reward function, $\hat{r}_t(s, a)$, is the output of the critic step. In Theorem 1 we show that the above procedure will result in a consistent $\hat{\mu}_t$. Thus for any $\epsilon > 0$, for large t, with high probability, $|\hat{r}_t(S, A)|_2$ will a.s. be less than $1 + \epsilon$. We point out that, when reward is bounded by a positive constant K other than 1, one shall modify the above projection by replacing 2 by $K + 1$.

Next the actor step maximizes the estimated $J_\lambda(\theta)$:

$$\hat{J}_t(\theta, \hat{\mu}_t) = \frac{1}{t} \sum_{\tau=1}^{t} \sum_a \hat{r}_t(S_\tau, a)\pi_\theta(S_\tau, a) - \lambda \theta^T \left(\frac{1}{t} \sum_{\tau=1}^{t} g(S_\tau)g(S_\tau)^T \right) \theta \qquad (7)$$

at decision point t to obtain $\hat{\theta}_t$. The action at decision point $t+1$ is selected according to the stochastic policy $\pi_{\hat{\theta}_t}(S_{t+1}, a)$. The actor critic algorithm, which alternates between a critic step and an actor step is depicted in Algorithm 1.

Algorithm 1: An online actor-critic algorithm with linear expected reward and stochastic policies

---

Inputs: T, the total number of decision points; a k dimensional reward feature $f(s, a)$; a p dimensional policy feature $g(s)$.

Critic initialization: $B(0) = \zeta I_{k \times k}$; $A(0) = \mathbf{0}_{k \times 1}$.

Actor initialization: θ_0 is initial policy parameter based on domain theory or historical data.

Start from $t = 0$.

while $t \leq T$ **do**

    At decision point t, observe context S_t.

    Draw an action A_t according to probability distribution $\pi_{\hat{\theta}_{t-1}}(S_t, A)$.

    Observe an immediate reward R_t.

    Critic update:

    $B(t) = B(t-1) + f(S_t, A_t)f(S_t, A_t)^T$, $A(t) = A(t-1) + f(S_t, A_t)R_t$,

    $\hat{\mu}_t = B(t)^{-1}A(t)$. The estimated reward function is $\hat{r}_t(s, a)$ from (6).

    Actor update:

$$\hat{\theta}_t = \underset{\theta}{\operatorname{argmax}} \frac{1}{t} \sum_{\tau=1}^{t} \sum_a \hat{r}_t(S_\tau, a)\pi_\theta(S_\tau, a) - \lambda \theta^T \left(\frac{1}{t} \sum_{\tau=1}^{t} g(S_\tau)g(S_\tau)^T \right) \theta.$$

    Go to decision point $t + 1$.

end

---

4 Asymptotic Theory for the Actor-Critic Algorithm

In this section, we present consistency and the asymptotic normality results for the proposed actor-critic algorithm. Proofs are provided in the supplementary material section C. In addition to assumptions 1, 2 and 3, we make the following assumption that ensures the identifiability of each component of the policy parameter.

Assumption 4. *(Positive definiteness of policy features) The $p \times p$ matrix $\mathbb{E}(g(S)g(S)^T) = \int_{s \in \mathcal{S}} d(s)g(s)g(s)^T ds$ is positive definite.*

As the very first step towards establishing the asymptotic properties of the actor-critic algorithm, we show that, for a fixed Lagrangian multiplier λ, the optimal policy parameter that maximizes the regularized average reward (3) lies in a bounded set. Moreover, the estimated optimal policy parameter is bounded with probability going to 1. Lemma 2 sets the foundation for us building on which we can use existing asymptotic statistical theory.

Lemma 2. *Assume that Assumption 3 and 4 hold. Given a fixed λ, the population optimal policy parameter, θ^*, lies in a compact set. In addition, the sequence of estimated optimal policy parameters, $\hat{\theta}_t$, lies in a compact set almost surely as $t \to \infty$. Denote this compact set by C_{θ^*}.*

Following this lemma, we make additional assumptions to establish the asymptotic theory of the actor-critic algorithm.

Assumption 5. *(Positive definiteness of reward features) The $k \times k$ matrix $\mathbb{E}_\theta(f(S, A)f(S, A)^T) = \int_{s \in \mathcal{S}} d(s) \sum_a f(s, a)f(s, a)^T \pi_\theta(s, a)ds$ is positive definite for all θ in the compact set C_{θ^*} in Lemma 2.*

Assumption 6. *(Uniform separateness of the global maximum) There exists a neighborhood of μ^*, say $B(\mu^*)$ such that the following holds. $J(\theta, \mu)$ as a function of θ has unique global maximum denoted by θ^μ, for each $\mu \in B(\mu^*)$. Moreover, for any $\delta > 0$, there exists $\epsilon > 0$ and neighborhood of θ^μ, denoted by $B(\theta^\mu, \epsilon)$, such that*

$$J(\theta^\mu, \mu) - \max_{\theta \notin B(\theta^\mu, \epsilon)} J(\theta, \mu) \geq \delta \tag{8}$$

for all $\mu \in B(\mu^*)$.

Under assumptions 1 through 6, the following theorems establish the consistency and asymptotic normality of the critic and the actor.

Theorem 1. *(Asymptotic properties of the critic) The $k \times 1$ vector $\hat{\mu}_t$ converges to the true reward parameter μ^* in probability. In addition, $\sqrt{t}(\hat{\mu}_t - \mu^*)$ converges in distribution to a multivariate normal with mean $\mathbf{0}_{k \times 1}$ and covariance matrix $[\mathbb{E}_{\theta^*}(f(S, A)f(S, A)^T)]^{-1}\sigma^2$, where $\mathbb{E}_\theta(f(S, A)f(S, A)^T) = \int_s d(s) \sum_a f(s, a)f(s, a)^T \pi_\theta(s, a)ds$ is the expected value of $f(S, A)f(S, A)^T$ under the policy with parameter θ, and σ is the standard deviation of the error term in Assumption 2. The plug-in estimator of the asymptotic covariance is consistent.*

Theorem 2. *(Asymptotic properties of the actor) The $p \times 1$ vector $\hat{\theta}_t$ converges to θ^* in probability. In addition, $\sqrt{t}(\hat{\theta}_t - \theta^*)$ converges in distribution to multivariate normal with mean $\mathbf{0}_{p \times 1}$ and covariance matrix $[J_{\theta\theta}(\mu^*, \theta^*]^{-1}V^*[J_{\theta\theta}(\mu^*, \theta^*)]^{-1}$, where*

$$V^* = \sigma^2 J_{\theta\mu}(\mu^*, \theta^*)\mathbb{E}_\theta[f(S, A)f(S, A)^T]J_{\mu\theta}(\mu^*, \theta^*) + \mathbb{E}[j_\theta(\mu^*, \theta^*, S)j_\theta(\mu^*, \theta^*, S)^T]$$

. In the expression of asymptotic covariance matrix,

$$j_\theta(\mu, \theta, S) = \frac{\partial}{\partial \theta}\left(\sum_a f(S, a)^T \mu \ \pi_\theta(S, a) - \lambda\theta^T[g(S)g(S)^T]\theta\right),$$

and both $J_{\theta\theta}$ and $J_{\theta\mu}$ are the second order partial derivatives with respect to θ twice and with respect θ and μ, respectively of J:

$$J(\mu, \theta) = \int_{s \in \mathcal{S}} d(s) \sum_{a \in \mathcal{A}} f(s, a)^T \mu \ \pi_\theta(s, a)ds - \lambda\theta^T\mathbb{E}[g(S)g(S)^T]\theta. \tag{9}$$

A bound on the expected regret can be derived as a by-product of the square-root convergence rate of $\hat{\theta}_t$. The expected regret of an online algorithm up to time T is the difference between the expected cumulative reward under the algorithm and that under the optimal policy θ^*:

$$U(T) = T\int_{s \in \mathcal{S}} d(s) \sum_a r(s, a)\pi_{\theta^*}(s, a)ds - \mathbb{E}\left[\sum_{t=1}^T R_t\right] \tag{10}$$

13

where $\{R_t\}_{t=1}^{T}$ is the sequence of rewards generated in the algorithm and the expectation E is with respect to the distribution of $\{S_t, A_t \sim \pi_{\hat{\theta}_{t-1}}, R_t\}_{t=1}^{T}$. Straightforward calculation shows that,

$$
\begin{aligned}
U(T) &= \sum_{t=1}^{T} \int_{s \in \mathcal{S}} d(s) \sum_{a} r(s,a)[\pi_{\theta^*}(s,a) - \mathbb{E}(\pi_{\hat{\theta}_{t-1}}(s,a))]ds \\
&= \sum_{t=1}^{T} \int_{s \in \mathcal{S}} d(s) \sum_{a} r(s,a)\mathbb{E}[\pi'_{\hat{\theta}_{t,s,a}}(s,a)(\theta^* - \hat{\theta}_{t-1})]ds
\end{aligned}
$$

where $\hat{\theta}_{t,s,a}$ is a random variable that lies on the line segment joining θ^* and $\hat{\theta}_{t-1}$. The boundedness of $r(s,a)$ together with Theorem 2 imply the following corollary.

Corollary 1. *The expected regret $U(T)$ of the actor-critic algorithm 1 is $O(\sqrt{T})$.*

Readers familiar with contextual bandit literature may wish to compare the above regret bound with the regret bounds for LinUCB (Chu et al. (2011)) and for Thompson sampling (Agrawal and Goyal (2013)). There are at least three differences between our framework and those considered in LinUCB and Thompson sampling papers. First, we parameterize explicitly both the policy class as well as the expected reward. These two papers parameterize the expected reward which then implicitly implies a parameterized deterministic policy class. As a result, our optimal policy is the policy that maximizes the regularized average reward over our explicitly defined policy class; however their optimal policy is the policy that maximizes the unregularized average reward. Second, we restrict our policy class to be stochastic whereas their implicitly defined policy class is composed of deterministic policies. Lastly, the setting considered here is more restrictive in that we assume contexts are i.i.d. whereas these papers allows for arbitrary contexts as long as the conditional mean of the reward in any context is linear in the context features.

5   Numerical Experiments

To assess the performance of the proposed actor-critic algorithm we conducted extensive simulations across a variety of realistic scenarios. Firstly, we conduct simulations to evaluate the relevance of our asymptotic theory in finite T settings in which the contexts are

indeed i.i.d. As will be seen, the bias and mean squared error (MSE) in estimating optimal policy decreases to 0 as sample size increases, and the bootstrap confidence interval for the optimal policy parameter achieves nominal confidence level. Secondly, we note that the i.i.d. assumption on the contexts is likely violated in real world applications in at least two ways: the current context may be influenced by the context at previous decision points and the current context may be influenced by past actions. Simulations in which the contexts follow an auto-regressive process show that the actor-critic algorithm is quite robust to auto-correlation among contexts. We also create simulation settings where context is influenced by previous actions through a burden effect of the treatments on the users. We observe reasonable robustness of the bandit actor-critic algorithm when the burden effects are small or moderate. Last but not least, we investigate how performance of the algorithm may deteriorate when assumption 2 is violated, that is, the conditional mean of the reward is non-linear.

Throughout we base the simulations on a generative model that is motivated by the Heartsteps application for improving daily physical activity (Klasnja et al. (2015); Dempsey et al. (2015)). A simplified description of HeartSteps follows. HeartSteps is a mobile health smartphone application seeking to reduce users' sedentary behavior and increase physical activity such as walking. A commercial wristband sensor is usted to collect minute level steps counts. Each evening self-report on the usefulness of the application as well as problems in daily life are collected. At each of 3 decision points per day, sensor data is collected including the user's current location (home/work/other) and weather. At each decision point, the algorithm on the smartphone application must decide whether to "push" a tailored physical activity suggestion, i.e., $A_t = 1$, or remain silent, i.e., $A_t = 0$. Our generative model uses a three dimensional context at decision point t: $S_t = [S_{t,1}, S_{t,2}, S_{t,3}]$. $S_{t,1}$ represents weather, with $S_{t,1} = -\infty$ being extremely severe and unfriendly weather for any outdoor activities and $S_{t,1} = \infty$ being the opposite. $S_{t,2}$ reflects the user's recent habits in engaging in physical activity. $S_{t,2} = \infty$ represents that the user has been maintaining positive daily physical habits while $S_{t,2} = -\infty$ represents the opposite. $S_{t,3}$ is a composite measure of disengagement with HeartSteps. $S_{t,3} = -\infty$ reflects an extreme state that the user is fully engaged, is adherent

and is reporting that the application is useful. On the other hand, $S_{t,3} = \infty$ denotes the opposite state of disengagement. Note that although we assumed bounded features in proving theoretical properties of the algorithm, these feature vectors have unbounded range. Results shown in later sections demonstrate robustness to the boundness assumption.

The goal of HeartSteps is to reduce users' sedentary behavior. Here we reverse code the reward and define the *cost* to be the sedentary time per hour between two decision points. So the goal of the actor-critic algorithm is to minimize an average penalized cost as opposed to maximizing an average penalized reward. The generative model for the cost is a linear model: $C_t = 10 - .4S_{t,1} - .4S_{t,2} - A_t \times (0.2 + 0.2S_{t,1} + 0.2S_{t,2}) + 0.4S_{t,3} + \xi_{t,0}$, where $\xi_{t,0}$ are i.i.d. N(0,1) errors. In this linear model, higher values of S_1 and S_2, good weather and positive physical activity habits, are associated with less sedentary time while a higher value of S_3, disengagement, leads to increased sedentary time. The negative main effect of A_t indicates that physical activity suggestion ($A_t = 1$) reduces sedentary behavior compared to no suggestion $A_t = 0$. The negative interaction between A_t and $S_{t,1}$ and between A_t and $S_{t,2}$ reflects that physical activity suggestions are more effective when the weather condition is activity friendly or when the user has acquired good physical activity habits.

The class of parametrized policies is $\pi_\theta(S, 1) = \frac{e^{\theta 0 + \sum_{i=1}^{3} \theta_i S_i}}{1 + e^{\theta 0 + \sum_{i=1}^{3} \theta_i S_i}}$. The average cost under policy π_θ is:

$$C(\theta) = \int_{s \in \mathcal{S}} d_\theta(s) \sum_a \mathbb{E}(C|S = a, A = a)\pi_\theta(s, a)ds$$

where $d_\theta(s)$ is the stationary distribution of context under policy π_θ. When actions have no impact on context distributions, the stationary distribution $d(s)$ does not depend on the policy parameter θ. In this case, the average cost reduces to: $C(\theta) = \int_{s \in \mathcal{S}} d(s) \sum_a \mathbb{E}(C|S = a, A = a)\pi_\theta(s, a)ds$. This is true for the generative models we investigate in Section 5.1 and Section 5.2. The generative model we investigate in Section 5.3 allows actions to impact the context distribution at future decision points. In such a case, the stationary distribution of context depends on the policy parameter θ. A quadratic constraint is enforced so that the optimal policy is stochastic. In the quadratic inequality (2), we use $\alpha = 0.1$ and $p_0 = 0.1$ throughout the numerical experiment unless otherwise specified. We then minimize the corresponding Lagrangian function.

The optimal policy θ^* and the oracle λ^*. According to results in Section 2.1, we have that for every pair of (p_0, α) there exists a Lagrangian multiplier λ^* such that the optimal solution to the regularized average cost function:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \, C(\theta) + \lambda \theta^T \sum_s d_\theta([1, s_1, s_2, s_3][1, s_1, s_2, s_3]^T)\theta \tag{11}$$

satisfies the quadratic constraint with equality. Furthermore, as λ increases the stringency of the quadratic constraint increases: an increased value of λ penalizes the quadratic term $\theta^{*T} \sum_s d_{\theta^*}([1, s_1, s_2, s_3][1, s_1, s_2, s_3]^T)\theta^*$ more heavily. For a fixed pair of (p_0, α), we perform a line search to find the smallest λ, denoted as λ^*, such that the minimizer to the regularized average cost, denoted as θ^* satisfies the quadratic constraint. We recognize the difficulty in solving the optimization problem due to the non-convexity of the regularized average cost function. In our search for a global minimizer, we therefore use grid search, for a given λ, to find a crude solution to the optimization problem. We then improve the accuracy of the optimal solution using a more refined grid search provided by the pattern search function in Matlab. The regularized average cost function is approximated by Monte Carlo samples. We used 5000 Monte Carlo samples to approximate the regularized average cost for simulation in Section 5.1 and Section 5.2 where the stationary distribution of contexts does not depend on the policy. For the simulations in Section 5.3, where context distribution does depend on the policy, we generate a trajectory of 100000 Monte Carlo samples and ignore the first 10% of the samples to approximate the stationary distribution.

Estimating λ online. In practice, the decision maker has no access to the oracle Lagrangian multiplier λ^*. A natural remedy is to integrate the estimation of λ^* with the online actor-critic algorithm that estimates the policy parameters. An actor-critic algorithm with a fixed Lagrangian multiplier solves the "primal" problem while the "dual" problem is solved by searching for λ^*. Our integrated algorithm performs a line search to find the smallest λ such that the estimated optimal policy satisfies the quadratic constraint. The stationary distribution of the contexts is approximated by the empirical distribution. Estimating λ can be very time consuming, therefore in our simulations, the algorithm performs the line search over λ only every 10 decision points. Similar ideas with gradient based updates on λ have appeared in reinforcement literature to find the optimal policies in constrained MDP

problems, see Borkar (2005); Bhatnagar and Lakshmanan (2012) for examples.

Bootstrap confidence intervals. In a number of trial simulations, we found that the plug-in variance estimator derived from Theorem 2 tends to underestimate in small to moderate sample size, a direct consequence of which is the anti-conservatism of the Wald confidence interval. Details of the anti-conservatism are discussed in the supplementary material section D. Our solution to the anti-conservative Wald confidence interval is the percentile-t bootstrap confidence interval. Algorithm 2 shows how to generate a bootstrap sample. Algorithm 2 is repeated for a total of B times to obtain a bootstrap sample of the estimated optimal policy parameters, $\{\hat{\theta}_T^b\}_{b=1}^B$ and plug-in variance estimates, $\{\hat{V}_T^b\}_{b=1}^B$. We create bootstrap percentile-t confidence intervals for θ_i^*, the i-th component of the optimal policy parameter. For each θ_i^*, we use the empirical percentile of $\left\{ \frac{\sqrt{t}(\hat{\theta}_{T,i}^b - \hat{\theta}_{T,i})}{\sqrt{\hat{V}_T^b}} \right\}_{b=1}^B$, denoted by p_α to replace the normal distribution percentile in Wald confidence intervals. A $(1-2\alpha)\%$ confidence interval is

$$\left[\hat{\theta}_{T,i} - p_\alpha \frac{\hat{V}_i}{\sqrt{T}}, \hat{\theta}_{T,i} + p_\alpha \frac{\hat{V}_i}{\sqrt{T}} \right] \tag{12}$$

where $\hat{\theta}_{T,i}$ is the i-th component of $\hat{\theta}_T$ and $\hat{V}_i$ is the plug-in variance estimate based on the original sample.

Simulation details. The simulation results presented in the following sections are based on 1000 independent simulated users. For each simulated user, we allow a burn-in period of 20 decision points. During the burn-in period, actions are chosen by fair coin flips. After the burn-in period, the online actor-critic algorithm is implemented to learn the optimal policy and obtain an end-of-study estimated optimal policy at the last decision point. In these simulations we did not force $\hat{r}_t(s, a)$ to be in the interval $[-2, 2]$ as in Algorithm 1 or Algorithm 2. We do not encounter any issues in convergence of the algorithm.

Both bias and MSE shown in all of the following tables are averaged over 1000 end-of-study estimated optimal policies. For each simulated user the 95% bootstrapped confidence intervals for θ^* is based on 500 bootstrapped samples generated by Algorithm 2. With 95% confidence, we expect that the empirical coverage rate of a confidence interval should be within 0.936 and 0.964, if the true confidence level is 0.95.

Algorithm 2: Generating a bootstrap sample estimate $\hat{\theta}_T^b, \hat{V}_T^b$

Inputs: The observed context history $\{S_t\}_{t=1}^T$. A bootstrap sample of residuals

$\{\epsilon_t^b\}_{t=1}^T$. The estimated reward parameter $\hat{\mu}_T$

Critic initialization: $B(0) = \zeta I_{k \times k}$, a $k \times k$ identity matrix. $A(0) = 0_k$ is a $k \times 1$

column vector.

Actor initialization: $\hat{\theta}_0^b = \hat{\theta}_0$ is the best treatment policy based on domain theory or

historical data.

while $t < T$ **do**

$\quad$ Context is S_t ;

$\quad$ Draw an action A_t^b according to policy $\pi_{\hat{\theta}_{t-1}^b}$;

$\quad$ Generate a bootstrap reward $R_t^b = f(S_t, A_t^b)^T \hat{\mu}_T + \epsilon_t^b$;

$\quad$ Critic update:

$\quad$ $B(t) = B(t-1) + f(S_t, A_t)f(S_t, A_t)^T$, $A(t) = A(t-1) + f(S_t, A_t)R_t^b$;

$\quad$ $\hat{\mu}_t^b = A(t)^{-1}B(t)$. The bounded estimate to reward function is $\hat{r}_t^b(s, a)$. ;

$\quad$ Actor update:

$$\hat{\theta}_t^b = \underset{\theta}{\operatorname{argmax}} \frac{1}{t} \sum_{\tau=1}^t \sum_a \hat{r}_t^b(S_\tau, a)\pi_\theta(a|S_t) - \lambda\theta^T[\frac{1}{t}\sum_{\tau=1}^t g(S_\tau, 1)^T g(S_\tau, 1)]\theta$$

$\quad$ Go to decision point $t+1$;

end

Plugin $\hat{\mu}_T^b$ and $\hat{\theta}_T^b$ to the asymptotic variance formula to get a bootstrapped variance

estimate $\hat{V}_T^b$.

T (sample size)	Bias				MSE			
	θ_0	θ_1	θ_2	θ_3	θ_0	θ_1	θ_2	θ_3
200	-0.081	-0.090	-0.089	0.010	0.054	0.052	0.052	0.055
500	-0.053	-0.037	-0.034	-0.002	0.027	0.024	0.021	0.029

Table 1: I.I.D. contexts: bias and MSE in estimating the optimal policy parameter. Bias=$\mathbb{E}(\hat{\theta}_T) - \theta^*$

.

T(sample size)	θ_0	θ_1	θ_2	θ_3
200	0.962	0.942	0.938	0.945
500	0.96	0.948	0.968	0.941

Table 2: I.I.D. contexts: coverage rates of percentile-t bootstrap confidence intervals for the optimal policy parameter.

5.1 I.I.D. Contexts

In this generative model, we choose the simplest setting where contexts at different decision points are i.i.d. We generate contexts $\{[S_{t,1}, S_{t,2}, S_{t,3}]\}_{t=1}^T$ from a multivariate normal distribution with mean 0 and identity covariance matrix. The population optimal policy is $\theta^* = [0.417778, 0.394811, 0.389474, 0.001068]$ at $\lambda^* = 0.046875$. Table 1 lists the bias and mean squared error (MSE) of the estimated optimal policy parameters. Both measures shrink towards 0 as T, sample size per simulated user, increases from 200 to 500, which is consistent with the convergence in estimated optimal policy parameter as established in Theorem 2. Table 2 shows the empirical coverage rates of percentile-t bootstrap confidence interval at sample sizes 200 and 500. At sample size 200, the empirical coverage rates are between 0.936 and 0.964 for all θ_i's. At sample size 500, however, the bootstrap confidence interval for θ_2 is a little conservative with an empirical coverage rate of 0.968.

20

5.2   AR(1) Context

In this section, we study the performance of the actor-critic algorithm when the dynamics of the context is an auto-regressive stochastic process. We envision that in many health applications, contexts at adjacent decision points are likely to be correlated. Using Heart-Steps as an example, weather (S_1) at two adjacent decisions points are likely to be similar. So are users' learning ability (S_2) and disengagement level S_3. One way to incorporate the correlation among contexts at near-by decision points is through a first order auto-regression process. We simulate the context according to

$$
\begin{aligned}
S_{t,1} &= 0.4S_{t-1,1} + \xi_{t,1}, \\
S_{t,2} &= 0.4S_{t-1,2} + \xi_{t,2}, \\
S_{t,3} &= \xi_{t,3}
\end{aligned}
$$

Here we choose $\xi_{t,1} \sim N(0, 1 - 0.4^2)$, $\xi_{t,2} \sim N(0, 1 - 0.4^2)$ and $\xi_{t,3} \sim N(0,1)$ so that the stationary distribution of S_t is multivariate normal with zero mean and identity covariance matrix, same as the distribution of S_t in the previous section. The initial distribution of $S_t, t = 1$ is a multivariate standard normal.

The oracle Lagrangian multiplier is $\lambda^* = 0.05$ and the population optimal policy is $\theta^* = [0.417, 0.395, 0.394, 0]$, same as in the i.i.d. simulation. Bias and MSE of the estimated policy parameters are shown in Table 3. Empirical coverage rate of the percentile t bootstrap confidence interval is reported in Table 4. Both the bias and MSE diminish towards 0 as the sample size increases from 200 to 500, a clear indication that convergence of the algorithm is not affected by the auto-correlation in context. The bootstrap confidence interval for θ_3 is anti-conservative at sample size 200, but recovers decent coverage at sample size 500.

5.3   Actions Cause Increased Burden

In this section, we study behavior of the actor-critic algorithm in the presence of an intervention burden effect. Our generative model with a burden effect represents a scenario where users disengage with the Heartsteps application, and hence the recommended intervention, if the application provides physical activity suggestions at too high a frequency. When users

T (sample size)	Bias				MSE			
	θ_0	θ_1	θ_2	θ_3	θ_0	θ_1	θ_2	θ_3
200	-0.093	-0.089	-0.076	0.006	0.058	0.053	0.047	0.057
500	-0.046	-0.032	-0.040	-0.005	0.025	0.022	0.024	0.028

Table 3: AR(1) contexts: bias and MSE in estimating the optimal policy parameter. Bias$=\mathbb{E}(\hat{\theta}_T) - \theta^*$.

T(sample size)	θ_0	θ_1	θ_2	θ_3
200	0.963	0.952	0.957	0.927*
500	0.969	0.962	0.96	0.949

Table 4: AR(1) contexts: coverage rates of percentile-t bootstrap confidence intervals. Coverage rates significantly lower than 0.95 are marked with asterisks (*).

experience intervention burden effects, they become frustrated and have a tendency of falling back to their sedentary behavior. In our burden effect generative model, $S_{t,3}$ represents the disengagement level whose value increases if there is a physical activity suggestion at the previous decision point $A_{t-1} = 1$. The positive main effect of $S_{t,3}$ in the cost model (13) below reflects that higher disengagement level is associated with higher cost (higher sedentary time). The initial distribution of S_t is the standard multivariate normal distribution. After the first decision point, contexts are generated according to the following stochastic process:

$$
\begin{aligned}
S_{t,1} &= 0.4S_{t-1,1} + \xi_{t,1}, \\
S_{t,2} &= 0.4S_{t-1,2} + \xi_{t,2}, \\
S_{t,3} &= 0.4S_{t-1,3} + 0.2S_{t-1,3}A_{t-1} + 0.4A_{t-1} + \xi_{t,3}
\end{aligned}
$$

We simulate the cost, sedentary time per hour between two decision points, according to the following linear model:

$$
C_t = \quad 10 - .4S_{t,1} - .4S_{t,2} - A_t \times (0.2 + 0.2S_{t,1} + 0.2S_{t,2}) + \tau S_{t,3} + \xi_{t,0}. \tag{13}
$$

where parameter τ controls the "size" of the burden effect: the larger τ is, the more severe the burden effect is. We study the performance of our algorithm in five different cases corresponding to $\tau = 0, 0.2, 0.4, 0.6, 0.8$. Different values of τ represent users who experience different levels of burden effect. $\tau = 0$ represents the type of users who experience no burden effect while $\tau = 0.8$ represents the type of users who experience a large burden effect.

Table 15 in the supplementary material section E lists the oracle λ^* and the corresponding optimal policy θ^* at different levels of burden effect. Higher level of burden effects calls for increased value of oracle λ^* to keep the desired intervention variety. The negative sign of θ_3^* at $\tau \geq 0.2$ indicates that the application should lower the probability of pushing an activity suggestion when the disengagement level is high. The magnitude of θ_3^* rises with the size of the burden effect, implying that as burden effect increases the application should further lower the probability of pushing activity suggestions at high disengagement level. θ_0^* decreases to be negative when τ increases, which indicates that as the size of burden effect grows, the application should lower the frequency of activity suggestions in general.

Table 5 and 6 list the bias, MSE and the empirical coverage rate of the percentile-t bootstrap confidence interval at sample size 200. Table 7 and 8 list these three measures at sample size 500. When there is no burden effect ($\tau = 0$), $S_{t,3}$ has no influence on the cost and is therefore considered as a "noise" variable. The optimal policy parameters are estimated with low bias and MSE under the generative model with $\tau = 0$ and the bootstrap confidence intervals have decent coverage, both of which are clear indications that the algorithm is robust to presence of noise variables that are affected by previous actions. As burden effects levels go up, we observe an increased bias and MSE in the estimated optimal policy parameters, θ_0 and θ_3 in particular. The empirical coverage rates of bootstrap confidence intervals for θ_0 and θ_3 are below the nominal 95% level. There are two reasons to explain the increased bias and MSE. The most important one is the near-sightedness of bandit actor-critic algorithm. The bandit algorithm chooses the policy that maximizes the (immediate) average cost while ignoring the negative consequence of a physical activity suggestion $A_t = 1$ on the disengagement level at the next decision point. The bandit algorithm therefore tends to "over-treat" in general and in particular at high disengagement level, which is reflected

in an over-estimated θ_0 and θ_3. The second reason comes from the bias in estimating λ, the Lagrangian multiplier. The oracle Lagrangian multiplier λ^* is chosen so that the optimal policy parameter satisfies the quadratic constraint while the online bandit actor-critic algorithm estimates the Lagrangian multiplier so that the bandit-estimated optimal policy satisfies the quadratic constraint. To separate the consequence of underestimated λ from the consequence of the myopia of the bandit algorithm, we implement the bandit algorithm with oracle λ^*. Results of these experiments are shown in the supplementary material section E. We observe that, even with the use of oracle λ^*, the overestimation of θ_0 and θ_3 as well as the anti-conservatism of the confidence intervals are still present.

Overall, the estimation of θ_1 and θ_2 shows robustness to the presence of burden effects. θ_1 and θ_2 are estimated with low bias and MSE under the presence of small to moderate burden effects ($\tau = 0.2, 0.4$). While we observe biases in estimating θ_1 and θ_2 under moderate to large burden effects ($\tau = 0.6, 0.8$), the magnitude of such bias increases slowly with the size of the burden effect. Empirical coverage rates of the bootstrap confidence intervals for θ_1 and θ_2 are decent for $\tau = 0.2, 0.4$ and only degrade slowly under 95% when $\tau = 0.6, 0.8$.

τ	Bias				MSE			
	θ_0	θ_1	θ_2	θ_3	θ_0	θ_1	θ_2	θ_3
0	-0.027	-0.036	-0.030	0.003	0.058	0.037	0.036	0.036
0.2	0.229	-0.093	-0.104	0.164	0.110	0.044	0.046	0.063
0.4	0.506	-0.063	-0.035	0.235	0.313	0.040	0.037	0.091
0.6	0.645	0.043	0.073	0.272	0.473	0.038	0.041	0.110
0.8	0.702	0.084	0.096	0.272	0.550	0.043	0.045	0.110

Table 5: Burden effect: bias and MSE in estimating the optimal policy parameter at sample size 200. Bias=$\mathbb{E}(\hat{\theta}_T) - \theta^*$.

τ	θ_0	θ_1	θ_2	θ_3
0	0.963	0.963	0.955	0.942
0.2	0.853*	0.946	0.937	0.862*
0.4	0.565*	0.96	0.954	0.776*
0.6	0.39*	0.937	0.916*	0.739*
0.8	0.329*	0.908*	0.899*	0.739*

Table 6: Burden effect: coverage rates of percentile-t bootstrap confidence intervals for the optimal policy parameter at sample size 200. λ is estimated online. Coverage rates significantly lower than 0.95 are marked with asterisks (*).
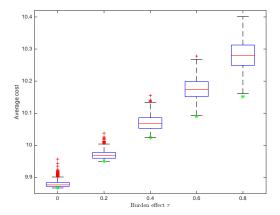
τ	Bias				MSE			
	θ_0	θ_1	θ_2	θ_3	θ_0	θ_1	θ_2	θ_3
0	0.006	0.010	0.017	-0.008	0.027	0.018	0.016	0.019
0.2	0.263	-0.048	-0.057	0.153	0.096	0.020	0.019	0.042
0.4	0.539	-0.018	0.012	0.224	0.318	0.018	0.016	0.069
0.6	0.678	0.088	0.120	0.261	0.487	0.026	0.030	0.087
0.8	0.735	0.129	0.143	0.261	0.568	0.035	0.037	0.087

Table 7: Burden effect: bias and MSE in estimating the optimal policy parameter at sample size 500. Bias=$\mathbb{E}(\hat{\theta}_t) - \theta^*$.

τ	θ_0	θ_1	θ_2	θ_3
0	0.973	0.949	0.955	0.942
0.2	0.714*	0.95	0.962	0.788*
0.4	0.217*	0.951	0.961	0.635*
0.6	0.101*	0.886*	0.835*	0.545*
0.8	0.07*	0.806*	0.788*	0.546*

Table 8: Burden effect: coverage rates of percentile-t bootstrap confidence intervals for the optimal policy parameter at sample size 200. λ is estimated online. Coverage rates significantly lower than 0.95 are marked with asterisks (*).

Figure 1 and 2 assess the quality of the estimated optimal policies by comparing the regularized average cost with the optimal regularized average cost. Figure 1 does the comparison at five levels of burden effect: $\tau = 0, 0.2, 0.4, 0.6, 0.8$, at sample size 200. As the burden effects level up, the overall long-run average cost goes up, which is simply an artifact of the increasing main effect size of the disengagement level. Having a higher long-term average cost, the estimated optimal policy by the contextual bandit algorithm is always inferior then the optimal policy. The inferiority gap, as measure by the difference between the median long-run average cost and the long-run average cost of the optimal policy increases as τ increases. When sample size increases from 200 to 500, we observe less variation in the long-run average cost of the estimated optimal policies. Nevertheless, the gap remains stable. We also observe that the variance in the regularized average cost increases as the burden effect level goes up.
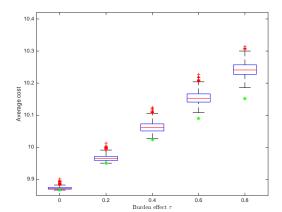
Figure 1: Burden effect: box plots of regularized average cost at different levels of the burden effect at sample size 200.

Figure 2: Burden effect: box plots of regularized average cost at different levels of the burden effect at sample size 500.

Because our theoretical results assume i.i.d. contexts, we have no proof that the optimal policy estimated by the bandit actor-critic algorithm will converge to the optimal policy. Nevertheless, we observe convergence in the estimated policy as sample size T grows. We conjecture that, when actions affect contexts distributions, the bandit algorithm converges to the policy $\pi_{\theta^{**}}$ that satisfies the following equilibrium equation:

$$\theta^{**} = \operatorname*{argmin}_{\theta} \sum_s d_{\theta^{**}}(s) \sum_a \pi_\theta(a|s)\mathbb{E}(C|A = a, S = s) - \lambda^{**}\theta^T \mathbb{E}_{\theta^{**}}[g(S)g(S)^T]\theta \quad (14)$$

where λ^{**} is the smallest λ such that $\theta^{**} \sum_s d_{\theta^{**}}(s)g(s)^T g(s)\theta^{**} \leq (\log(\frac{p_0}{1 - p_0}))^2\alpha \quad (15)$

When actions do not influence contexts distributions, the equilibrium equation is the same system of equations satisfied by the optimal policy. When previous actions have an impact on context distribution at later decision points, the stationary distribution of context is a function of policy. We call solution to equation 15, the *myopic equilibrium policy*. The myopic equilibrium policy minimizes the regularized average cost under the stationary distribution generated by itself. Such policy achieves an "equilibrium state" and there is no reason for the actor-critic to change the current policy if a myopic equilibrium has been reached.

27

The conjecture is supported by our numerical results. Since myopic equilibrium policy only depends on the context dynamics and the treatment effect $\mathbb{E}(C|A = 1, S = s) - \mathbb{E}(C|A = 0, S = s)$, it remains the same at different levels of the burden effect. The myopic equilibrium policy is $\theta^{**} = [0.392, 0.372, 0.371, 0]$. The bias and MSE in estimating the myopic equilibrium policy for $\tau = 0.4$ is shown in table 9. The bias and MSE at other levels of the burden effect are the same. These results support with our conjecture that the estimated optimal policy by the bandit algorithm converges to the myopic equilibrium policy.

Sample size (T)	Bias				MSE			
	θ_0	θ_1	θ_2	θ_3	θ_0	θ_1	θ_2	θ_3
200	-0.078	-0.081	-0.075	0.004	0.063	0.042	0.041	0.036
500	-0.045	-0.035	-0.028	-0.007	0.029	0.019	0.017	0.019

Table 9: Burden effect: bias and MSE in estimating the myopic equilibrium policy for $\tau = 0.4$. Bias=$\mathbb{E}(\hat{\theta}_t) - \theta^{**}$.

5.4 Expected Cost is a Nonlinear function of the Cost Feature

In this section, we investigate the performance of the online actor critic algorithm when the expected cost is a nonlinear function of the cost feature used in the critic step. In such scenarios, the linear actor critic algorithm finds the "best" policy in two steps: first it projects the true cost function into the linear space spanned by the cost feature, then it finds the policy that minimizes the regularized cost function under the projection. In contrast, the true optimal policy is the policy that minimizes the regularized cost function without the projection. In this simulation, we are interested to see how the extra step of projection affects the estimation and inference of the optimal policy parameter.

Recall that the cost feature is $f(S_t, A_t) = [1, S_{t,1}, S_{t,2}, S_{t,3}, A_t, A_t S_{t,1}, A_t S_{t,2}, A_t S_{t,3}]$. In particular consider the case where the interaction term between A_t and $S_{t,1}$ is a linear

combination of a linear cost function and a nonlinear one:

$$C_t = (1-\alpha)[10 - .4S_{t,1} - .4S_{t,2} - A_t \times (0.2 + 0.2S_{t,1} + 0.2S_{t,2}) + 0.4S_{t,3} + \xi_{t,0}]$$
$$+ \alpha[10 - .4S_{t,1}^2 - .4S_{t,2} - A_t \times (0.2 + 0.2S_{t,1}^2 + 0.2S_{t,2}) + 0.4S_{t,3} + \xi_{t,0}]$$
$$= 10 - .4[(1-\alpha)S_{t,1} + \alpha S_{t,1}] - .4S_{t,2} - A_t \times (0.2 + 0.2[(1-\alpha)S_{t,1} + \alpha S_{t,1}] + 0.2S_{t,2}) + 0.4S_{t,3} + \xi_{t,0}$$

The tuning parameter $\alpha \in [0,1]$ controls the amount of nonlinearity: when $\alpha = 0$, the expected cost is the linear cost function used in the previous sections. Nonlinearity increasingly dominates the interaction between $S_{t,1}$ and A_t as α increases. The online actor critic algorithm, unaware of the possible nonlinearity in the cost function, uses the same cost feature and the same policy feature as in the previous sections. Recall the policy is parameterized as $\pi_\theta(S,1) = \frac{e^{\theta_0 + \sum_{i=1}^3 \theta_i S_i}}{1 + e^{\theta_0 + \sum_{i=1}^3 \theta_i S_i}}$. Table 23 in the supplementary material section F provides the optimal θ values. Table 10 and Table 11 show the bias and MSE of the linear actor critic algorithm at different levels of nonlinearity at sample size 200 and 500. The bias for estimating θ_i^*, $i = 1, 2, 3$ remains stable whereas the MSE inflates as the α increases. Both the bias and MSE for estimating θ_0^* increase as the cost function moves away from a linear structure. Table 12 and table 13 show the coverage rates of the confidence interval for θ^*. The confidence interval coverages for θ_i^*, $i = 0.1, 2$ deteriorate as the level of nonlinearity increases. However, the confidence level for $\theta_3^* = 0$, the coefficient for $S_{t,3}$ which is not a useful tailoring variable, remains decent as the level of nonlinearity increases.

α	Bias				MSE			
	θ_0	θ_1	θ_2	θ_3	θ_0	θ_1	θ_2	θ_3
0	-0.100	-0.074	-0.103	-0.007	0.063	0.052	0.051	0.057
0.2	-0.137	-0.012	-0.109	-0.013	0.064	0.065	0.050	0.053
0.4	-0.179	0.012	-0.109	-0.014	0.076	0.099	0.048	0.048
0.6	-0.211	0.020	-0.099	-0.017	0.083	0.142	0.043	0.042

Table 10: Nonlinear Cost: bias and MSE in estimating the optimal policy parameter at sample size 200. Bias=$\mathbb{E}(\hat{\theta}_T) - \theta^*$.

α	Bias				MSE			
	θ_0	θ_1	θ_2	θ_3	θ_0	θ_1	θ_2	θ_3
0	-0.038	-0.036	-0.045	-0.002	0.025	0.022	0.022	0.027
0.2	-0.080	0.030	-0.067	-0.008	0.026	0.032	0.023	0.023
0.4	-0.109	0.065	-0.069	-0.010	0.031	0.064	0.023	0.020
0.6	-0.136	0.058	-0.068	-0.009	0.038	0.112	0.021	0.018

Table 11: Nonlinear Cost: bias and MSE in estimating the optimal policy parameter at sample size 500. Bias=$\mathbb{E}(\hat{\theta}_T) - \theta^*$.

α	θ_0	θ_1	θ_2	θ_3
0	0.944	0.947	0.954	0.939
0.2	0.926*	0.879*	0.942	0.935*
0.4	0.892*	0.738*	0.922*	0.942
0.6	0.835*	0.588*	0.914*	0.942

Table 12: Nonlinear Cost: coverage rates of percentile-t bootstrap confidence intervals for the optimal policy parameter at sample size 200. λ is estimated online. Coverage rates significantly lower than 0.95 are marked with asterisks (*).

α	θ_0	θ_1	θ_2	θ_3
0	0.971	0.961	0.966	0.958
0.2	0.931*	0.875*	0.936	0.956
0.4	0.885*	0.655*	0.924*	0.958
0.6	0.837*	0.471*	0.915*	0.961

Table 13: Nonlinear Cost: coverage rates of percentile-t bootstrap confidence intervals for the optimal policy parameter at sample size 500. λ is estimated online. Coverage rates significantly lower than 0.95 are marked with asterisks (*).

6 Conclusion and Discussion

In this article, we present a general framework to define optimal policies for use in JITAIs that encourages intervention variety. We also gave an online actor-critic algorithm to learn the optimal policy. Although the theoretical properties of the algorithm assume i.i.d. contexts, the numerical experiments show robustness of the algorithm to violations of this assumption. In particular, experiments show that performance of the algorithm, in term of bias, MSE and confidence interval coverage, is not affected by auto-correlation among contexts. Experiments also demonstrate some robustness of the algorithm when distribution of the context depends on previous actions. Furthermore, we conjecture that, when actions influence the distribution of context at later decision points, the contextual bandit algorithm converges to the myopic equilibrium policy. Our numerical experiments back up this conjecture. Theoretical proof of the conjecture, however, is an open question and requires future work.

There are a few areas for which the actor-critic algorithm could be improved and extended. First, the linear expected reward assumption might be a bit strong in some scenarios, especially when a low dimension reward feature is used. When the assumption is deemed untenable, more sophisticated components should be added to the reward (cost) feature. To this end, both the actor-critic algorithm and the asymptotic theory should be extended to encompass the scenario where the dimension of the reward (cost) feature grows with the sample size. If one intends to use linear reward (cost) model with a fixed dimension of reward feature, we highly recommend frequent validation of the linear model using model diagnostic tools. Linear regression diagnostic tools can be used as the first line of defense. However, more sophisticated model checking methods for online learning need to be developed to make sure the reward (cost) model is adequate. Second, there is room for improvement in optimization in the actor step. Optimizing the estimated regularized average reward function is in general a non-convex optimization problem and could be time-consuming. In the proposed algorithm, optimization at decision point $t + 1$ does not use the estimated policy parameters at previous decision points. In other words, the optimization is not incremental and may waste computing resources when the sample size gets large. Careful design of online optimization methods that leverages previous estimates will likely

significantly improve the computational efficiency of the actor-critic algorithm and help in its practical adoption in mobile health applications. Third, the algorithm presented in this article learns a user's the optimal policy based *solely* on his/her history. However, in order to speed up the learning it is attractive idea, especially in the beginning of the learning period, to pool data across multiple users. Methods and theories for learning based on multiple users need to be developed.

References

Agrawal, S. and N. Goyal (2013). Thompson sampling for contextual bandits with linear payoffs. In *ICML (3)*, pp. 127–135.

Bauer, S., J. de Niet, R. Timman, and H. Kordy (2010). Enhancement of care through self-monitoring and tailored feedback via text messaging and their use in the treatment of childhood overweight. *Patient education and counseling 79*(3), 315–319.

Bertsekas, D. P. (1999). Nonlinear programming.

Bhatnagar, S. and K. Lakshmanan (2012). An online actor–critic algorithm with function approximation for constrained markov decision processes. *Journal of Optimization Theory and Applications 153*(3), 688–708.

Billingsley, P. (1961). The lindeberg-levy theorem for martingales. *Proceedings of the American Mathematical Society 12*(5), 788–792.

Borkar, V. S. (2005). An actor-critic algorithm for constrained markov decision processes. *Systems & control letters 54*(3), 207–213.

Campi, M. C. and S. Garatti (2011). A sampling-and-discarding approach to chance-constrained optimization: feasibility and optimality. *Journal of Optimization Theory and Applications 148*(2), 257–280.

Chu, W., L. Li, L. Reyzin, and R. E. Schapire (2011). Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, pp. 208–214.

Consolvo, S., D. W. McDonald, T. Toscos, M. Y. Chen, J. Froehlich, B. Harrison, P. Klasnja, A. LaMarca, L. LeGrand, R. Libby, et al. (2008). Activity sensing in the wild: a field trial of ubifit garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1797–1806. ACM.

Dempsey, W., P. Liao, P. Klasnja, I. Nahum-Shani, and S. A. Murphy (2015). Randomised trials for the fitbit generation. *Significance 12*(6), 20–23.

Epstein, L. H., K. A. Carr, M. D. Cavanaugh, R. A. Paluch, and M. E. Bouton (2011). Long-term habituation to food in obese and nonobese women. *The American journal of clinical nutrition 94*(2), 371–376.

Epstein, L. H., J. L. Temple, J. N. Roemmich, and M. E. Bouton (2009). Habituation as a determinant of human food intake. *Psychological review 116*(2), 384.

Fiacco, A. V. and Y. Ishizuka (1990). Sensitivity and stability analysis for nonlinear programming. *Annals of Operations Research 27*(1), 215–235.

Gustafson, D. H., B. R. Shaw, A. Isham, T. Baker, M. G. Boyle, and M. Levy (2011). Explicating an evidence-based, theoretically informed, mobile technology-based system to improve outcomes for people in recovery for alcohol dependence. *Substance use & misuse 46*(1), 96–111.

Keener, R. (2010). *Theoretical Statistics: Topics for a Core Course*. Springer Texts in Statistics. Springer New York.

King, A. C., E. B. Hekler, L. A. Grieco, S. J. Winter, J. L. Sheats, M. P. Buman, B. Banerjee, T. N. Robinson, and J. Cirimele (2013). Harnessing different motivational frames via mobile phones to promote daily physical activity and reduce sedentary behavior in aging adults. *PloS one 8*(4), e62613.

Klasnja, P., E. B. Hekler, S. Shiffman, A. Boruvka, D. Almirall, A. Tewari, and S. A. Murphy (2015). Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology 34* (Suppl), 1220–1228.

Langford, J. and T. Zhang (2008). The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pp. 817–824.

Li, L., W. Chu, J. Langford, and R. E. Schapire (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670. ACM.

Nahum-Shani, I., S. N. Smith, A. Tewari, K. Witkiewitz, L. M. Collins, B. Spring, and S. Murphy (2014). Just in time adaptive interventions (jitais): An organizing framework for ongoing health behavior support. *Methodology Center technical report* (14-126).

Nemirovski, A. and A. Shapiro (2006). Convex approximations of chance constrained programs. *SIAM Journal on Optimization 17* (4), 969–996.

Patrick, K., F. Raab, M. Adams, L. Dillon, M. Zabinski, C. Rock, W. Griswold, and G. Norman (2009). A text message-based intervention for weight loss: randomized controlled trial. *Journal of medical Internet research 11* (1), e1.

Prékopa, A. (1995). *Stochastic programming.* Springer Science & Business Media.

Raynor, H. A. and L. H. Epstein (2001). Dietary variety, energy regulation, and obesity. *Psychological bulletin 127* (3), 325.

Riley, W. T., D. E. Rivera, A. A. Atienza, W. Nilsen, S. M. Allison, and R. Mermelstein (2011). Health behavior models in the age of mobile interventions: are our theories up to the task? *Translational behavioral medicine 1* (1), 53–71.

Scott, C. K. and M. L. Dennis (2009). Results from two randomized clinical trials evaluating the impact of quarterly recovery management checkups with adult chronic substance users. *Addiction 104* (6), 959–971.

Suffoletto, B., C. Callaway, J. Kristan, K. Kraemer, and D. B. Clark (2012). Text-message-based drinking assessments and brief interventions for young adults discharged from the emergency department. *Alcoholism: Clinical and Experimental Research 36*(3), 552–560.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics 12*(4), 389–434.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

Wilson, T. D., D. B. Centerbar, D. A. Kermer, and D. T. Gilbert (2005). The pleasures of uncertainty: prolonging positive moods in ways people do not anticipate. *Journal of personality and social psychology 88*(1), 5.

Witkiewitz, K., S. A. Desai, S. Bowen, B. C. Leigh, M. Kirouac, and M. E. Larimer (2014). Development and evaluation of a mobile intervention for heavy drinking and smoking among college students. *Psychology of Addictive Behaviors 28*(3), 639.

Woodroofe, M. (1979). A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association 74*(368), 799–806.

Supplementary Material

A Proof of Lemma 1

Proof. Without the loss of generality, we assume that $0 < s_1 < s_2 < ... < s_K$. Otherwise, if some s_i's are negative, we can transform all the contexts to be positive by adding to s_i's a constant greater than $\min_{1 \le i \le K} s_i$. Denote this constant by M and the corresponding policy parameter by $\tilde{\theta}$. There is a one-to-one correspondence between the two policy classes:

$$\tilde{\theta}_0 = \theta_0 - M\theta_1$$

$$\tilde{\theta}_1 = \theta_1$$

Therefore if the lemma holds when all contexts are positive the same conclusion hold in the general setting. We use $p(\theta)$ to denote the probability the probability of choosing action $A = 1$ for policy π_θ at the K different values of context:

$$\left(\frac{e^{\theta_0 + \theta_1 s_1}}{1 + e^{\theta_0 + \theta_1 s_1}}, \frac{e^{\theta_0 + \theta_1 s_2}}{1 + e^{\theta_0 + \theta_1 s_2}}, ..., \frac{e^{\theta_0 + \theta_1 s_K}}{1 + e^{\theta_0 + \theta_1 s_K}} \right)$$

Notice that each entry in $p(\theta)$ is number between 0 and 1 with equality if the policy is deterministic at certain context. A key step towards proving deterministic optimal policy is to show the following closed convex hull equivalency:

$$conv(\{p(\theta) : \theta \in \mathbb{R}^2\}) = conv(\{(\nu_1, ..., \nu_K), \nu_i \in \{0, 1\}, \nu_1 \le ... \le \nu_K \text{ or } \nu_1 \ge ... \ge \nu_K\})$$

We examine the limiting points of $p(\theta)$ when θ_0 and θ_1 tends to infinity. We consider the case where $\theta_0 \ne 0$ and let $\theta_1 = p\theta_0$ where p is a fixed value. It holds that

$$\frac{e^{\theta_0 + \theta_1 s}}{1 + e^{\theta_0 + \theta_1 s}} = \frac{e^{\theta_0(1+ps)}}{1 + e^{\theta_0(1+ps)}} \rightarrow \begin{cases} 0 : if \theta_0 \rightarrow -\infty, p > -1/s \\ 0 : if \theta_0 \rightarrow \infty, p < -1/s \\ 1 : if \theta_0 \rightarrow -\infty, p < -1/s \\ 1 : if \theta_0 \rightarrow \infty, p > -1/s \end{cases}$$

It follows that when $\theta_0 \rightarrow -\infty$ and p scans through the $K + 1$ intervals on $\mathbb{R}$: $(-\infty, -1/s_1]$, $(-1/s_1, -1/s_2]$, ... $(-1/s_K, \infty)$, $p(\theta)$ approaches the following $K + 1$ limiting points:

$$(1, 1, ..., 1)$$

$$(0, 1, ..., 1)$$

$$...$$

$$(0, 0, ..., 1)$$

$$(0, 0, ..., 0)$$

when $\theta_0 \to \infty$ and p scans through the $K + 1$ intervals, $p(\theta)$ approaches the following $K + 1$ limiting points

$$(0, 0, ..., 0)$$

$$(1, 0, ..., 0)$$

$$...$$

$$(1, 1, ..., 0)$$

$$(1, 1, ..., 1)$$

There are in total $2K$ limiting points: $\{(\nu_1, ..., \nu_K), \nu_i \in \{0, 1\}, \nu_1 \leq ... \leq \nu_K \text{ or } \nu_1 \geq ... \geq \nu_K\}$. Each limiting point is a K dimensional vector with 0-1 entries in an either increasing or decreasing order. Now we show that any $p(\theta), \theta \in \mathbb{R}^2$ is a convex combination of the limiting points. Let $p(\theta) = [p_1(\theta), p_2(\theta), ..., p_K(\theta)]$. In fact,

- If $\theta_1 = 0$, $p(\theta) = (1 - p_1(\theta))(0, 0, ..., 0) + p_1(\theta)(1, 1, ..., 1)$

- If $\theta_1 > 0$, we have $0 < p_1(\theta) < p_2(\theta) < ... < p_K(\theta) < 1$ and

$$p(\theta) = p_1(\theta)(1, 1, ..., 1) + (p_2(\theta) - p_1(\theta))(0, 1, ..., 1) + ...$$
$$+ (p_K(\theta) - p_{K-1}(\theta))(0, 0, ..., 1) + (1 - p_K(\theta)) * (0, 0, ..., 0)$$

- If $\theta_1 < 0$, we have $1 > p_1(\theta) > p_2(\theta) > ... > p_K(\theta) > 0$ and

$$p(\theta) = (1 - p_1(\theta)) * (0, 0, ..., 0) + (p_1(\theta) - p_2(\theta))(1, 0, ..., 0) + ...$$
$$+ (p_K(\theta) - p_{K-1}(\theta))(1, 1, ..., 0) + p_K(\theta)(1, 1, ..., 1)$$

Returning to optimizing the average reward, we denote $\alpha_i = P(S = s_i)(\mathbb{E}(R|S = s_i, A = 1) - \mathbb{E}(R|S = s_i, A = 0))$.

$$\max_\theta V^*(\theta) = \max_\theta \sum_{i=1}^K \alpha_i p_i(\theta) \tag{16}$$

$$= \max_{(p_1,...,p_K)\in\{p(\theta):\theta\in\mathbb{R}^2\}} \sum_{i=1}^K \alpha_i p_i \tag{17}$$

$$= \max_{(p_1,...,p_K)\in conv(\{p(\theta):\theta\in\mathbb{R}^2\})} \sum_{i=1}^K \alpha_i p_i \tag{18}$$

$$= \max_{(p_1,...,p_K)\in conv(\{(\nu_1,...,\nu_K),\nu_i\in\{0,1\},\nu_1\leq...\leq\nu_K \text{ or } \nu_1\geq...\geq\nu_K\})} \sum_{i=1}^K \alpha_i p_i \tag{19}$$

. Equation from (17) to (18) is followed by the fact that the objective function is linear (and thus convex) in p_i's. Equivalency from (18) to (19) is a direct product of the closed convex hull equivalency. Theories in linear programming theory suggests that one of the maximal points is attained at the vertices of the convex hull of the feasible set. Therefore we have proved that one of the policy that maximizes $V^*(\theta)$ is deterministic. $\qquad\square$

B One-to-one Correspondence between Constrained and Unconstrained Optimization

The constrained optimization finds the policy that maximizes the average reward subject to the quadratic constraint, i.e.,

$$\max_\theta V^*(\theta), \text{ s. t. } \theta^T \mathbb{E}[g(S)^T g(S)]\theta \leq (\log(\frac{p_0}{1 - p_0}))^2 \alpha \tag{20}$$

The unconstrained optimization finds the policy that maximizes the regularized average reward:

$$\theta^* = \underset{\theta}{\operatorname{argmax}} J_\lambda^*(\theta) \tag{21}$$

A natural question to ask, when transforming the constrained optimization problem (20) to an unconstrained one (21), is whether a Lagrangian multiplier exists for each level of

stringency of the quadratic constraint. While the correspondence between the constrained optimization and the unconstrained one may not seem so obvious due to the lack of convexity in $V^*(\theta)$, we established the following lemma 3 given assumption 7 and assumption 4. Assumption 7 assumes the uniqueness of the global maximum for all positive λ.

Assumption 7. *For every $0 < \lambda < \infty$, the global maximum of the regularized average reward is a singleton.*

$$J_\lambda^*(\theta) = \sum_{s \in \mathcal{S}} d(s) \sum_{a \in \mathcal{A}} E(R|S = s, A = a)\pi_\theta(s, a) - \lambda\theta^T \mathbb{E}[g(S)^T g(S)]\theta$$

Lemma 3. *If the maximizer of the average reward function $V^*(\theta)$ is deterministic, i.e. $P(\pi_\theta(A = 1|S) = 1) > 0$ or $P(\pi_\theta(A = 0|S) = 1) > 0$, under assumption 7 and 4, for every $K = (\log(\frac{p_0}{1-p_0}))^2\alpha > 0$ there exist a $\lambda > 0$ such that the solution of the constrained optimization problem 20 is the solution of the unconstrained optimization problem 21.*

Proof. Let θ_λ^* be one of the global maxima of the Lagrangian function: $\theta_\lambda^* = \text{argmax}_\theta J_\lambda^*(\theta)$. Let $\beta_\lambda = \theta_\lambda^{*T}\mathbb{E}[g(S)^T g(S)]\theta_\lambda^*$. By proposition 3.3.4 in Bertsekas (1999), θ_λ^* is a global maximum of constrained problem:

$$\max_\theta V^*(\theta)$$

$$\text{s.t. } \theta^T \mathbb{E}[g(S)^T g(S)]\theta \le \beta_\lambda$$

In addition, the stringency of the quadratic constraint increases monotonically with the value of the Lagrangian coefficient λ. Let $0 < \lambda_1 < \lambda_2$ and with some abuse of notation, let θ_1 and θ_2 be (one of) the global maximals of Lagrangian function $J_{\lambda_1}^*(\theta)$ and $J_{\lambda_2}^*(\theta)$. It follows that

$$- V^*(\theta_2) + \lambda_2\theta_2^T \mathbb{E}[g(S)^T g(S)]\theta_2$$
$$\le - V^*(\theta_1) + \lambda_2\theta_1^T \mathbb{E}[g(S)^T g(S)]\theta_1$$
$$= - V^*(\theta_1) + \lambda_1\theta_1^T \mathbb{E}[g(S)^T g(S)]\theta_1 + (\lambda_2 - \lambda_1)\theta_1^T \mathbb{E}[g(S)^T g(S)]\theta_1$$
$$\le - V^*(\theta_2) + \lambda_1\theta_2^T \mathbb{E}[g(S)^T g(S)]\theta_2 + (\lambda_2 - \lambda_1)\theta_1^T \mathbb{E}[g(S)^T g(S)]\theta_1$$

It follows that

$$\theta_1^T \mathbb{E}[g(S)^T g(S)]\theta_1 \geq \theta_2^T \mathbb{E}[g(S)^T g(S)]\theta_2$$

. As λ approaches 0, the maximal of the regularized average reward approaches the maximal of the average reward function, for which $\mathbb{E}(\theta^T g(S))^2 \to \infty$. As λ increases towards ∞, maximal of the regularized average reward approaches the random policy with $\theta = 0$. It's only left to show that $\theta_\lambda^{*T} \mathbb{E}[g(S)^T g(S)]\theta_\lambda^*$ is a continuous function of λ. Under assumption 7, we can verify that conditions in Theorem 2.2 in Fiacco and Ishizuka (1990) holds. This theorem implies that the solution set of the unconstrained optimization 21 is continuous in λ, sufficient to conclude the continuity of $\theta_\lambda^{*T} \mathbb{E}[g(S)^T g(S)]\theta_\lambda^*$. □

C Proof of Asymptotic Theory of the Actor Critic Algorithm

C.1 Proof of Lemma 2

Proof. This lemma is proved by comparing the regularized average reward function $J_\lambda^*(\theta)$ at θ^* and at 0_p. The optimal regularized average reward is:

$$\begin{aligned}
J_\lambda^*(\theta^*) &= \sum_{s \in \mathcal{S}} d(s) \sum_{a \in \mathcal{A}} f(s,a)^T \mu^* \pi_{\theta^*}(A = a | S = s) - \lambda \theta^{*T} \mathbb{E}[g(S)^T g(S)]\theta^* \\
&\leq \sum_{s,a} d(s) \frac{|f(s,a)|_2^2 + |\mu^*|_2^2}{2} \pi_{\theta^*}(A = a | S = s) - \lambda \theta^{*T} \mathbb{E}[g(S)^T g(S)]\theta^* \\
&\leq 1 - \lambda \theta^{*T} \mathbb{E}[g(S)^T g(S)]\theta^*
\end{aligned}$$

While on the other hand the regularized average reward for the random policy $\theta = 0_p$ is

$$J_\lambda^*(0_p) = \sum_{s,a} d(s) f(s,a)^T \mu^* / 2 \geq 0$$

By the optimality of policy θ^*, $1 - \lambda \theta^T \mathbb{E}[g(S)^T g(S)]\theta \geq 0$, which leads to the necessary condition for the optimal policy parameter:

$$\theta^{*T} \mathbb{E}[g(S)^T g(S)]\theta^* \leq \frac{1}{\lambda} \tag{22}$$

According to assumption 4, the above inequality defines a bounded ellipsoid for θ^*, which concludes the first part of the lemma. To prove the second part of this lemma, we notice that the estimated reward function $\hat{r}_t(s, a)$ is by definition bounded. By comparing $\hat{J}_t(\theta, \hat{\mu}_t)$ at $\theta = \hat{\theta}_t$ and $\theta = 0_p$ we have

$$\hat{\theta}_t^T [\frac{1}{t} \sum_{\tau=1}^{t} g(S_\tau)g(S_\tau)^T]\hat{\theta}_t \leq \frac{2}{\lambda} \tag{23}$$

It remains to show that the smallest eigenvalue of $\frac{1}{t}\sum_{\tau=1}^{t} g(S_\tau)g(S_\tau)^T$ is bounded away from 0 with probability going to 1. Using the matrix Chernoff inequality, theorem 1 in Tropp (2012), for any $0 < \delta < 1$,

$$P\{\lambda_{min}(\frac{1}{t}\sum_{\tau=1}^{t} g(S_\tau)g(S_\tau)^T) \leq (1-\delta)\lambda_{min}(\mathbb{E}g(S)g(S)^T)\} \leq p[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}]^{\frac{t\lambda_{min}(\mathbb{E}g(S)g(S)^T)}{Q}} \tag{24}$$

where Q is the bound on the maximal eigenvalue of $g(S)g(S)^T$ and p is the length of $g(S)$. Taking $\delta = 0.5$, the right-hand side of the Chernoff inequality goes to 0 as t goes to ∞. Therefore with probability going to 1, inequality 23 defines a compact set on $\mathbb{R}^p$. We have proved the second part of the lemma.

$\square$

C.2 Proof of the consistency of the critic

Proof. Based on the expression of $\hat{\mu}_t$, its $\mathcal{L}_2$ distance from μ^* is

$$|\hat{\mu}_t - \mu^*|^2 = C(t)B(t)^{-1}B(t)^{-1}C(t) + o_p(1) \tag{25}$$

$$= \frac{C(t)}{t}(\frac{B(t)}{t})^{-1}(\frac{B(t)}{t})^{-1}\frac{C(t)}{t} + o_p(1) \tag{26}$$

where

$$C(t) = \sum_{\tau=1}^{t} f(S_\tau, A_\tau)\epsilon_\tau$$

$$B(t) = \zeta I_{k \times k} + \sum_{\tau=1}^{t} f(S_\tau, A_\tau)f(S_\tau, A_\tau)^T$$

The two steps in proving $|\hat{\mu}_t - \mu^*|_2^2 \to 0$ in probability are

1. The matrix $\frac{B(t)}{t}$ has eigenvalue bounded away from 0 with probability going to 1, and

2. $\frac{C(t)}{t}$ converges to 0_d in probability.

To prove the first step, we construct a matrix-valued martingale difference sequence. Define
$$K(\theta) = \mathbb{E}_\theta[f(S,A)f(S,A)^T] = \sum_s d(s) \sum_a f(s,a)f(s,a)^T \pi_\theta(A = a|S = s)$$

$$
\begin{aligned}
X_i &= f(S_i, A_i)f(S_i, A_i)^T - \mathbb{E}(f(S_i, A_i)f(S_i, A_i)^T | \mathcal{F}_i) \\
&= f(S_i, A_i)f(S_i, A_i)^T - \int_s d(s) \sum_a f(s,a)f(s,a)^T \pi_{\theta_{i-1}}(s,a)ds \\
&= f(s_i, a_i)f(s_i, a_i)^T - K(\hat{\theta}_{i-1})
\end{aligned}
$$

where the filtration $\mathcal{F}_i = \sigma\{\hat{\theta}_j, j \leq i - 1\}$ is the sigma algebra expand by the estimated optimal policy before decision point i. By assumption 3, the sequence of random matrices $\{X_i\}$ are uniformly bounded. Applying the matrix Azuma inequality in Tropp (2012), it follows that

$$\lambda_{max}\left(\frac{B(t)}{t} - \frac{\sum_{i=1}^{t} K(\hat{\theta}_{i-1})}{t}\right) \to 0 \text{ in probability}$$

$$\lambda_{min}\left(\frac{B(t)}{t} - \frac{\sum_{i=1}^{t} K(\hat{\theta}_{i-1})}{t}\right) \to 0 \text{ in probability}$$

Let the operators λ_{min} and λ_{max} represent the smallest and the largest eigenvalue of a matrix.

$$
\begin{aligned}
\lambda_{min}\left(\frac{B(t)}{t}\right) &= \lambda_{min}\left(\frac{B(t)}{t} - \frac{\sum_{i=1}^{t} K(\hat{\theta}_{i-1})}{t} + \frac{\sum_{i=1}^{t} K(\hat{\theta}_{i-1})}{t}\right) \\
&\geq \lambda_{min}\left(\frac{B(t)}{t} - \frac{\sum_{i=1}^{t} K(\hat{\theta}_{i-1})}{t}\right) + \lambda_{min}\left(\frac{\sum_{i=1}^{t} K(\hat{\theta}_{i-1})}{t}\right)
\end{aligned}
$$

By assumption 5, the second term $\lambda_{min}(\frac{\sum_{i=1}^t K(\hat{\theta}_{i-1})}{t})$ is bounded with probability going to 1. Hence we have shown that the minimal eigenvalue of $\frac{B(t)}{t}$ is bounded with probability going to 1. Using the same proving techniques we can show that the maximal eigenvalue of $(\frac{B(t)}{t})^{-1}$ is bounded with probability going to 1.

The second step in proving consistency of the critic is standard. Using the same filtration $\mathcal{F}_i$, we construct vector-valued martingale difference sequence $Y_i = f(S_i, A_i)\epsilon_i$. The sequence has bounded variance under assumption 3. The in-probability convergence of $\frac{C(t)}{t}$ to 0 follows immediately by applying the vector-valued Azuma inequality.

$\square$

C.3 Proof of the consistency of the actor

First we prove the following lemma, which will be utilized in proving both the consistency and the asymptotic normality of the actor.

Lemma 4. *Define $\tilde{\theta}_t$ to be the estimated optimal policy parameter by plugging in the possible unbounded estimated reward function:*

$$\tilde{\theta}_t = \underset{\theta}{\operatorname{argmax}} \frac{1}{t} \sum_{\tau=1}^t \sum_a f(S_\tau, a)^T \hat{\mu}_t \pi_\theta(S_\tau, a) - \lambda \theta^T \left(\frac{1}{t} \sum_{\tau=1}^t g(S_\tau) g(S_\tau)^T \right) \theta.$$

The probability that $\tilde{\theta}_t \neq \hat{\theta}_t$ goes to 0 as $t \to \infty$.

Proof. The proof starts by noticing that $\hat{\theta}_t$ and $\tilde{\theta}_t$ are equal if $|f(s,a)^T \hat{\mu}_t|$ are bounded by 2 for all combinations of $(S_\tau, a), 1 \leq \tau \leq t, a \in \mathcal{A}$.

$$
\begin{aligned}
P(\tilde{\theta}_t \neq \hat{\theta}_t) &\leq P(\exists 1 \leq \tau \leq t, a \in \mathcal{A}, s.t. |f(S_\tau, a)^T \hat{\mu}_t| > 2) \\
&\leq P(\exists 1 \leq \tau \leq t, a \in \mathcal{A}, \text{ s.t. } |f(S_\tau, a)^T(\hat{\mu}_t - \mu^*)| + |f(S_\tau, a)^T \mu^*| > 2) \\
&\leq P(\exists 1 \leq \tau \leq t, a \in \mathcal{A}, \text{ s.t. } |f(S_\tau, a)^T(\hat{\mu}_t - \mu^*)| > 1) \\
&\leq P(\exists 1 \leq \tau \leq t, a \in \mathcal{A}, \text{ s.t. } |f(S_\tau, a)|_2 |\hat{\mu}_t - \mu^*|_2 > 1) \\
&\leq P(|\hat{\mu}_t - \mu^*|_2 > 1) \\
&\to 0
\end{aligned}
$$

where the third inequality is based on $|f(S_\tau, a)^T \mu^*| \leq 1$ a.s. from assumption 3. The fourth inequality is based on Holder's inequality. In the end we use the consistency of $\hat{\mu}_t$. □

Now we can prove the consistency of the actor by proving the consistency of $\tilde{\theta}_t$.

Proof. Proof of the theorem consists of two steps. As the first step, we claim that if a sequence μ_t converges to μ^*, $\tilde{\theta}_t = \text{argmax}_\theta J(\theta, \mu_t)$ converges to θ^*. By Lemma 9.1 in Keener (2010), $J(\theta, \mu)$ is an absolute continuous function. We proof the claim by contradiction. Suppose the claim does not hold, i.e. there exist ϵ such that $\|\tilde{\theta}_t - \theta^*\|_2 \geq \epsilon$ for all t by taking a subsequence if necessary. The optimality of $\tilde{\theta}_t$ implies that the inequality $J(\tilde{\theta}_t, \mu_t) \geq J(\theta^*, \mu_t)$ holds for all t. Since $\tilde{\theta}_t$ is bounded, it converges to an accumulation point $\tilde{\theta}$ by taking a subsequence if necessary. Let $t \to \infty$ we have $J(\tilde{\theta}, \mu^*) \geq J(\theta^*, \mu^*)$. On the other hand $\|\theta^{**} - \theta^*\|_2 \geq \epsilon$, which contradicts with assumption 7. As the second step, we prove that the following M-estimator converges uniformly in a neighborhood of μ^*, namely

$$\theta_t^\mu = \text{argmax}_\theta \frac{1}{t} \sum_{\tau=1}^{t} \sum_{a \in \mathcal{A}} f(S_\tau, a)^T \mu \pi_\theta(S_\tau, a) - \theta^T [\frac{1}{t} \sum_{\tau=1}^{t} g(S_\tau)g(S_\tau)^T] \theta$$

$$\to \theta^\mu = \text{argmax}_\theta J(\theta, \mu)$$

in probability, and uniformly over all μ in a neighborhood of μ^*. Arguments in the proof are parallel to those in Theorem 9.4 in Keener (2010). The key is to observe that the class of random functions $\{j(\theta, \mu, s) = \sum_{a \in \mathcal{A}} f(s, a)^T \mu \pi_\theta(s, a) - \theta^T g(s)g(s)^T \theta : \theta \in \mathbb{R}^p, |\mu|_2 \leq 1\}$ are Glivenko-Cantelli. We have now proved the consistency of $\tilde{\theta}_t$ and thus the consistency of $\hat{\theta}_t$. □

C.4   Proof of the asymptotic normality of the critic

Proof. Based on the formula of $\hat{\mu}_t$,

$$\hat{\mu}_t - \mu^* = (\zeta I_d + \sum_{i=1}^{t} f(S_i, A_i)f(S_i, A_i)^T)^{-1}(\sum_{i=1}^{t} f(S_i, A_i)\epsilon_i - \mu^*)$$

$$= (\frac{\zeta I_d + \sum_{i=1}^{t} f(S_i, A_i)f(S_i, A_i)^T}{t})^{-1}\sqrt{t}\frac{\sum_{i=1}^{t} f(S_i, A_i)\epsilon_i}{t} + o_p(1)$$

Based on the consistency of θ_t, we have that $\frac{\zeta I_d + \sum_{i=1}^t f(S_i, A_i) f(S_i, A_i)^T}{t}$ converges in probability to $\mathbb{E}_{\theta^*}(f(S, A) f(S, A)^T)$. Now it is the key to analyze the asymptotic distribution of the martingale difference sequence $\{f(S_i, A_i)\epsilon_i\}_{i=1}^t$. With respect to filtration $\mathcal{F}_{t,j} = \sigma(\{S_i, A_i, \epsilon_i\}_{i=1}^j)$. Define $M* = [\mathbb{E}_{\theta^*}(f(S, A) f(S, A)^T)]^{-1/2}$ and a martingale difference sequence $\{\xi_{t,i} = \frac{M^* f(s_i, a_i)\epsilon_i}{\sqrt{t}}\}_{i=1}^t$ which is adapted to the filtration $\mathcal{F}_{t,j}$ and satisfies $\mathbb{E}(\xi_{t,i}|\mathcal{F}_{t,i-1}) = 0$, To apply vector Lindberg-Levy central limit theorem for martingale difference sequences (Billingsley (1961)), we check the two conditions in this theorem:

1. The conditional variance assumption.

$$V_t = \sum_{i=1}^t \mathbb{E}(\xi_{t,i}^2 | \mathcal{F}_{t,i-1})$$

$$= \frac{1}{t} \sum_{i=1}^t M^* \mathbb{E}_{\theta_{i-1}}(f(s,a) f(s,a)^T) M^*$$

   converges in probability to $I_d \sigma^2$ by consistency of θ_t.

2. The Lindeberg condition. For any given $\delta > 0$,

$$\sum_{i=1}^t \mathbb{E}(\xi_{t,i}^2 \mathbb{I}(\|\xi_{t,i}\|_2 > \delta) | \mathcal{F}_{t,i-1})$$

$$= \frac{1}{t} \sum_{i=1}^t \mathbb{E}(M^* f(S_i, A_i) f(S_i, A_i)^T \epsilon_i^2 M^* \mathbb{I}(\|M^* f(S_i, A_i)\epsilon_i\|_1 > \sqrt{t}\delta) | \mathcal{F}_{t,i-1})$$

$$\leq \frac{1}{t} \sum_{i=1}^t \mathbb{E}(M^* f(S_i, A_i) f(S_i, A_i)^T \epsilon_i^2 M^* \mathbb{I}(\|M^* f(S_i, A_i)\|_2 \epsilon_i^2 > \sqrt{t}\delta) | \mathcal{F}_{t,i-1})$$

   By assumption 3, $f(S, A)$ are bounded almost surely, therefore the above expression goes to 0 as $t \to 0$.

The Lindberg-Levy martingale central limit theorem concludes that

$$\sum_{i=1}^t \xi_{t,i} \to N(0_d, I_d \sigma^2) \text{ in distribution}$$

Therefore

$$\sqrt{t}(\hat{\mu}_t - \mu^*) \to N(0_d, [\mathbb{E}_{\theta^*}(f(S, A) f(S, A)^T)]^{-1}\sigma^2) \tag{27}$$

$\square$

C.5 Proof of the asymptotic normality of the actor

Again, our strategy is to derive the asymptotic normality of $\tilde{\theta}_t$ and then use the fact that $\hat{\theta}_t$ must have the same asymptotic distribution.

Proof. We first prove that

$$\mathbb{G}_t j_\theta(\hat{\mu}_t, \hat{\theta}_t, S) - \mathbb{G}_t j_\theta(\mu^*, \theta^*, S) = o_p(1) \tag{28}$$

, where $\mathbb{G}_t = \sqrt{t}(\mathbb{P}_t - P)$, the empirical process induced by the "marginal" stochastic process $\{S_i\}_{i=1}^t$ formed by the history of contexts. The "full" stochastic process involves the sequence of triples $\{S_i, A_i, \epsilon_i\}_{i=1}^t$, the complete history of contexts, actions and reward errors. We consider the class of functions $\mathcal{F} = \{j_\theta(\mu, \theta, s) : \|\theta - \theta^*\|_2 \leq \delta, \|\mu - \mu^*\|_2 \leq \delta\}$, where $j_\theta(\mu, \theta, s)$ is the partial derivative with respect to θ of function:

$$j(\mu, \theta, s) = \sum_a f(s, a)^T \mu \pi_\theta(s, a) - \lambda \theta^T g(s) g(s)^T \theta$$

The boundedness assumption on reward feature, policy feature and reward ensures that the parametrized class of functions $j_\theta(\mu, \theta, s)$ is P-Donsker in a neighborhood of (μ^*, θ^*). In other words $\mathcal{F}$ is P-Donsker, where P is the distribution of the marginal stochastic process formed by contexts. We complete the first part of the proof by modiftying Lemma 19.24 in Van der Vaart (2000). It may seem that the dependence of $\hat{\mu}_t$ and $\tilde{\theta}_t$ on the full stochastic process could introduce complexity but a closer inspection shows that the proof goes through. The random function $j_\theta(\hat{\mu}_t, \tilde{\theta}_t, S)$ belongs to the P-Donsker class defined above and satisfies that

$$\sum_s d(s)(j_\theta(\hat{\mu}_t, \tilde{\theta}_t, s) - j_\theta(\mu^*, \theta^*, s))^2 \to 0$$

in probability. This is a result of the consistency of both $\hat{\mu}_t$ and $\tilde{\theta}_t$, as well as applying the continuous mapping theorem. By Theorem 18.10(v) in Van der Vaart (2000), $(\mathbb{G}_t, j_\theta(\hat{\mu}_t, \tilde{\theta}_t, s)) \to (\mathbb{G}_p, j_\theta(\mu^*, \theta^*, s))$ in distribution, where $\mathbb{G}_p$ is the P-Brownian bridge. The key here is that Theorem 18.10 only relies on the convergence of two stochastic processes, regardlessly of

whether the stochastic processes consist of i.i.d. observations and whether or not the two processes are dependent. By Lemma 18.15 in Van der Vaart (2000), almost all sample paths of $\mathbb{G}_p$ are continuous on $\mathcal{F}$. Define a mapping $h : l(\mathcal{F})^\infty \times \mathcal{F} \to \mathbb{R}$ by $h(z, f) = z(f) - z(j_\theta(\mu^*, \theta^*, s))$, which is continuous at almost every point of $(\mathbb{G}_p, j_\theta(\mu^*, \theta^*, s))$. By the continuous mapping theorem, we have

$$\mathbb{G}_t(j_\theta(\hat{\mu}_t, \tilde{\theta}_t, s) - j_\theta(\mu^*, \theta^*, s)) = h(\mathbb{G}_t, j_\theta(\hat{\mu}_t, \tilde{\theta}_t, s)) \to h(\mathbb{G}_p, j_\theta(\mu^*, \theta^*, s)) = 0$$

in distribution and thus in probability, therefore (28) holds. The second part of the proof begins by noticing that $\tilde{\theta}_t$ satisfies the estimating equation $\mathbb{P}_t j_\theta(\hat{\mu}_t, \tilde{\theta}_t, s) = 0$, so we have

$$
\begin{aligned}
\mathbb{G}_t j_\theta(\hat{\mu}_t, \tilde{\theta}_t, s) &= \sqrt{t}(P j_\theta(\mu^*, \theta^*, s) - P j_\theta(\hat{\mu}_t, \tilde{\theta}_t, s)) \\
&= \sqrt{t}(J_\theta(\mu^*, \theta^*) - J_\theta(\hat{\mu}_t, \tilde{\theta}_t)) \\
&= \sqrt{t} J_{\theta\theta}^*(\theta^* - \tilde{\theta}_t) + \sqrt{t} J_{\theta\mu}^*(\mu^* - \hat{\mu}_t) + \sqrt{t} o_p(\|\tilde{\theta}_t - \theta^*\|) + o_p(1)
\end{aligned}
$$

Together with (28) the above implies

$$
\begin{aligned}
\sqrt{t}(\theta^* - \tilde{\theta}_t) &= (J_{\theta\theta}^*)^{-1} J_{\theta\mu}^* \sqrt{t}(\hat{\mu}_t - \mu^*) + \sqrt{t} o_p(\|\tilde{\theta}_t - \theta^*\|) + (J_{\theta\theta}^*)^{-1} \mathbb{G}_t j_\theta(\mu^*, \theta^*, s) + o_p(1) \\
&= O_p(1) + \sqrt{t} o_p(\|\tilde{\theta}_t - \theta^*\|)
\end{aligned}
$$

where $J_{\theta\theta}^*$ and $J_{\theta\mu}^*$ are $J_{\theta\theta}$ and $J_{\theta\mu}$ evaluated at (θ^*, μ^*). The $\sqrt{t}$ consistency of $\tilde{\theta}_t$ follows through. Now (29) has become

$$\sqrt{t}(\theta^* - \tilde{\theta}_t) = (J_{\theta\theta}^*)^{-1} J_{\theta\mu}^* \sqrt{t}(\hat{\mu}_t - \mu^*) + (J_{\theta\theta}^*)^{-1} \mathbb{G}_t j_\theta(\mu^*, \theta^*, S) + o_p(1) \tag{29}$$

Since both the two non-vanishing terms on the righthand side are asymptotically normal with zero mean, $\sqrt{t}(\theta^* - \tilde{\theta}_t)$ is asymptotically normal. The only task left is to derive the asymptotic variance. Plugging in the formula for $\hat{\mu}_t$, we have

$$
\begin{aligned}
\sqrt{t}(\theta^* - \tilde{\theta}_t) &= (J_{\theta\theta}^*)^{-1} \frac{\sum_{i=1}^t J_{\theta\mu}^* B^* f(S_i, A_i) \epsilon_i + j_\theta(\mu^*, \theta^*, S_i)}{t} + o_p(1) \\
&= (J_{\theta\theta}^*)^{-1} \sum_{i=1}^t \zeta_{t,i} + o_p(1)
\end{aligned}
$$

A12

where $B^* = (M^*)^2 = [\mathbb{E}_{\theta^*}f(S,A)f(S,A)^T]^{-1}$. $\{\zeta_i = \frac{J^*_{\theta\mu}B^*f(S_i,A_i)\epsilon_i + j_\theta(\mu^*,\theta^*,S_i)}{t}\}_{i=1}^t$ is a martin-gale difference sequence with asymptotic variance

$$\sum_{i=1}^t \mathbb{E}(\zeta_{t,i}^2|\mathcal{F}_{t,i})$$

$$= \frac{1}{t}\sum_{i=1}^t \mathbb{E}(\epsilon_i^2 g^*_{\theta\mu}B^*f(S_i,A_i)f(S_i,A_i)^T B^* g^*_{\mu\theta}$$

$$+ j_\theta(\mu^*,\theta^*,S_i)j_\theta(\mu^*,\theta^*,S_i)^T - 2J_{\theta\mu}B^*f(S_i,A_i)j_\theta(\mu^*,\theta^*,S_i)^T\epsilon_i|\mathcal{F}_{t,i})$$

$$= \frac{1}{t}\sum_{i=1}^t \sigma^2 J^*_{\theta\mu}B^*\mathbb{E}_{\theta_{i-1}}(f(S,A)f(S,A)^T)B^* J^*_{\mu\theta} + \sum_s d(s)j_\theta(\mu^*,\theta^*,s)j_\theta(\mu^*,\theta^*,s)^T$$

which converges in probability to $V^* = \sigma^2 J^*_{\theta\mu}B^* J^*_{\mu\theta} + \sum_s d(s)j_\theta(\mu^*,\theta^*,s)j_\theta(\mu^*,\theta^*,s)^T$. There-fore the asymptotic variance of $\sqrt{t}(\theta^* - \tilde{\theta}_t)$ is $(J^*_{\theta\theta})^{-1}V^*(J^*_{\theta\theta})^{-1}$.

$\square$

D   Small Sample Variance estimation and Bootstrap Confidence intervals

In this section, we discuss issues, challenges and solutions in creating confidence intervals for the optimal policy parameter θ^* when the sample size, the total number of decision points, is small. We use a simple example to illustrate that the traditional plug-in variance estimator is plagued with underestimation issue, the direct consequence of which is the deflated confidence levels of the Wald-type confidence intervals for θ^*. We propose to use bootstrap confidence intervals when the sample size is finite. We use simulation to evaluate the bootstrap confidence intervals.

D.1   Plug-in Variance Estimation and Wald Confidence intervals

One of the most straightforward ways to estimate the asymptotic variance of θ_t is through the plug-in variance estimation, the formulae of which is provided in Theorem 2. Once an

estimated variance $\hat{V}_i$ is obtained for $\sqrt{t}(\hat{\theta}_i - \theta_i^*)$, a $(1 - 2\alpha)\%$ Wald type confidence interval for θ_i^* has the form: $[\hat{\theta}_i - z_\alpha \frac{\hat{V}_i}{\sqrt{t}}, \hat{\theta}_i + z_\alpha \frac{\hat{V}_i}{\sqrt{t}}]$. Here θ_i is the i-th component in θ and z_α is the upper 100α percentile of a standard normal distribution. The plug-in variance estimator and the associated Wald confidence intervals work well in many statistics problems. We shall see that, however, the plug-in variance estimator of the estimated optimal policy parameters suffers from underestimation issue in small to moderate sample sizes. In particular this estimator is very sensitive to the plugged-in value of the estimated reward parameter and policy parameter: a small deviation from the true parameters can result in an inflated or deflated variance estimation. Deflated variance estimation produces anti-conservative confidence intervals, a grossly undesirable property for confidence intervals. The following simple example illustrates the problem.

Example 1. *The context is binary with probability distribution $\mathbb{P}(S = 1) = \mathbb{P}(S = -1) = 0.5$. The reward is generated according to the following linear model: given context $S \in \{-1, 1\}$ and action $A \in \{0, 1\}$,*

$$R = \mu_0^* + \mu_1^* S + \mu_2^* A + \mu_3^* SA + \epsilon$$

where ϵ follows a normal distribution with mean zero and standard deviation 9. The true reward parameter is $\mu^* = [1, 1, 1, 1]$. Both μ^* and the standard deviation of ϵ are chosen to approximate the realistic signal noise ratio in mobile health applications. We consider the policy class $\pi_\theta(A = 1 | S = s) = \frac{e^{\theta_0 + \theta_1 s}}{1 + e^{\theta_0 + \theta_1 s}}$.

The differences between the plug-in estimated variance and its population counterpart are that (1) the former uses the empirical distribution of context to replace the unknown population distribution and (2) the unknown reward parameter and optimal policy parameter are replaced by their estimates. We emphasize that it is the second difference that leads to the underestimated variance in small sample size. To see this, we ignore the difference between the empirical distribution and the population distribution of contexts, which is very small for sample size $T \geq 50$ under a Bernoulli context distribution with equal probability. Now the plug-in variance estimator is a function of the estimated reward parameter $\hat{\mu}_t$ and the estimated policy parameter $\hat{\theta}_t$. Notice that $\hat{\theta}_t = [\hat{\theta}_{t,0}, \hat{\theta}_{t,1}]$ is a function of $\hat{\mu}_t = [\hat{\mu}_{t,0}, \hat{\mu}_{t,1}, \hat{\mu}_{t,2}, \hat{\mu}_{t,3}]$

A14

and the empirical distribution of context. If we replace the empirical distribution in calcu-

lating $\hat{\theta}_t$ by its population counterpart, $\hat{\theta}_t$ is simply a function of $\hat{\mu}_t$. In the rest part of the

example, we drop the subscript t in the estimated reward parameter and denote the estimate

of μ_2 and μ_3 by $\hat{\mu}_2$ and $\hat{\mu}_3$, respectively. Likewise, $\hat{\theta}_{t,i}$ is replaced by $\hat{\theta}_i$ for $i = 0, 1$.

   Figure 3 is the surface plot showing how the plug-in variance estimation changes as func-

tion of the estimated reward parameter. The surface plot of the plug-in variance estima-

tion has a mountain-like pattern with two ridges along the two diagonals $\hat{\mu}_2 + \hat{\mu}_3 = 0$ and

$\hat{\mu}_2 - \hat{\mu}_3 = 0$. The height of the ridge increases as both $\hat{\mu}_2$ and $\hat{\mu}_3$ approaches the origin.

The peak of mountain is at the origin where $\hat{\mu}_2 = \hat{\mu}_3 = 0$. The true reward parameter

$(\mu_2^*, \mu_3^*) = (1, 1)$ is close to the origin and lies right on the one of the ridges. There are four

"valleys" where the combinations of $\hat{\mu}_2$ and $\hat{\mu}_3$ gives a small plug-in variance.

Figure 3: Plug in variance estimation as a function of
$\hat{\mu}_2$ and $\hat{\mu}_3$, x axis represents $\hat{\mu}_{t,2}$, y axis represents $\hat{\mu}_{t,3}$ and z axis represents the plug-in
asymptotic variance of $\hat{\theta}_0$ with $\lambda = 0.1$

   Due to large areas of valley the plug-in variance estimation is biased down, a direct

consequence of which is the anti-conservatism of the Wald confidence intervals. We perform

a simulation study using the toy generative model described above. The simulation consists

of 1000 repetitions of running the online actor critic algorithm and recording the end-of-

study statistics, including the plugin variance estimate, the Wald confidence intervals and the theoretical Wald confidence intervals based on the true asymptotic variance. The first two columns in table 14 show the bias of plug-in variance at different sample sizes. At all three different sample sizes, the plug-in variance estimator underestimates the true asymptotic variance, which is 293.03 for both policy parameters. Column 3 and column 4 show the coverage rate of the Wald-type confidence interval (CI) using the plug-in estimated variance. It is not surprising that the confidence intervals suffer from severe anti-conservatism, a consequence of the heavily biased variance estimation. Column 5 and 6 show the coverage rate of the Wald-type confidence interval based on the true asymptotic variance. Comparing the coverage rates, it is clear that the anti-conservatism is due to the underestimated variance.

sample size	bias in variance estimation		coverage of Wald CI (%)		coverage of theoretical Wald CI (%)	
	θ_0	θ_1	θ_0	θ_1	θ_0	θ_1
100	-181.56	-181.56	75.5	74.9	100.0	100.0
250	-131.71	-131.71	77.9	77.3	98.5	98.1
500	-108.64	-108.64	78.8	79.2	98.9	98.7

Table 14: Underestimation of the plug-in variance estimator and the Wald confidence intervals.

Theoretical Wald CI is created based on the true asymptotic variance.

To detail how the confidence interval coverage is connected with the estimated reward parameter $(\hat{\mu}_2, \hat{\mu}_3)$, figure 4 and figure 5 present two scatter plots of $\hat{\mu}_2, \hat{\mu}_3$ for the 1000 simulated datasets at sample size 100 and 500. Different colors are used to mark the datasets where the confidence intervals of both θ_0 and θ_1 cover the true parameter (blue), only one of them cover the truth (green), neither of them covers the truth (fading yellow). The true parameter are marked with a red asterisk. Indeed the yellow points and green points are in the "valleys". Some of the blue points are away from truth, but nevertheless they remain on the ridge, which produces a high variance estimate. Comparing the two scatter plots, as the sample size increases, the estimated reward parameter is less spread out. Nevertheless there are still significantly many pair of $\hat{\mu}_2, \hat{\mu}_3$ that fall in the "valleys", leading to a underestimated variance and anti-conservative confidence intervals.
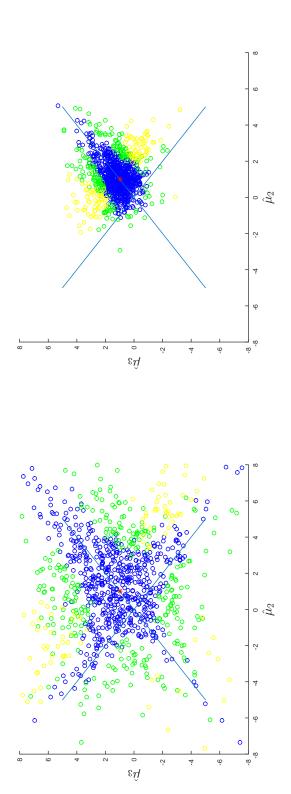
Figure 4: Wald confidence interval coverage for 1000 simu-Figure 5: Wald confidence interval coverage in 1000 simulated datasets as a function of $\hat{\mu}_3$ and $\hat{\mu}_2$ at sample size 100.datasets as a function of $\hat{\mu}_3$ and $\hat{\mu}_2$ at sample size 500.

A19

Figure 6 shows the histogram for the normalized distance $\frac{\sqrt{\hat{T}}(\theta_i - \theta_i^*)}{\hat{V}_i}$ for $i = 0, 1$ where $T = 100$. This is the distance between the estimated and the true optimal policy parameter normalized by the estimated asymptotic variance. For the Wald confidence intervals to have descent coverage, histogram of the normalized distances need to approximate a standard normal distribution. However, as figure 6 suggests, the histograms have heavier tails compared to a standard normal due to the underestimated variance. The figure also suggests that the percentile-t bootstrap confidence intervals can be a good remedy.
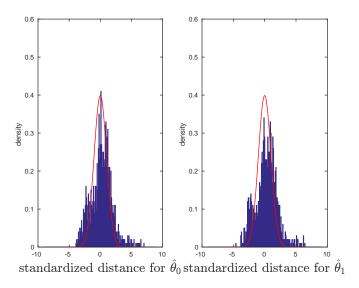

Figure 6: Histograms of the normalized distance $\frac{\sqrt{\hat{T}}(\theta_i - \theta_i^*)}{\hat{V}_i}$ for $i = 0, 1$ at sample size 100

E   Burden Effect: Actor Critic Algorithm Uses λ^*

ν	λ^*	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0.0	0.06	0.3410	0.3269	0.3264	0
0.2	0.05	0.0844	0.3844	0.4	-0.1609
0.4	0.06	-0.1922	0.3547	0.3312	-0.2313
0.6	0.08	-0.3312	0.2488	0.2234	-0.2687
0.8	0.1	-0.3883	0.2078	0.2	-0.2687

Table 15: Burden effect: the optimal policy and the oracle lambda.

τ	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	$-0.027\,352$	$-0.035\,565$	$-0.030\,344$	$0.003\,449$
0.2	$0.229\,47$	$-0.092\,877$	$-0.104\,06$	$0.164\,21$
0.4	$0.505\,86$	$-0.063\,199$	$-0.035\,223$	$0.234\,73$
0.6	$0.645\,07$	$0.042\,695$	$0.072\,542$	$0.271\,98$
0.8	$0.702\,29$	$0.083\,867$	$0.096\,08$	0.2718

Table 16: Burden effect: bias in estimating the optimal policy parameter at sample size 200. The algorithm uses λ^* instead of learning λ online. Bias=$\mathbb{E}(\hat{\theta}_t) - \theta^*$.

τ	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	0.057 811	0.037 16	0.036 343	0.035 898
0.2	0.109 61	0.044 463	0.046 192	0.062 836
0.4	0.312 95	0.039 819	0.036 65	0.090 984
0.6	0.473 09	0.037 714	0.040 625	0.109 84
0.8	0.550 24	0.042 799	0.044 54	0.1097

Table 17: Burden effect: MSE in estimating the optimal policy parameter at sample size 200. The algorithm uses λ^* instead of learning λ online.

ν	θ_0	θ_1	θ_2	θ_3
0	0.963	0.963	0.955	0.942
0.2	0.853*	0.946	0.937	0.862*
0.4	0.565*	0.96	0.954	0.776*
0.6	0.39*	0.937	0.916*	0.739*
0.8	0.329*	0.908*	0.899*	0.739*

Table 18: Burden effect: coverage rates of percentile-t bootstrap confidence intervals for the optimal policy parameter at sample size 200. The algorithm uses λ^* instead of learning λ online. Coverage rates significantly lower than 0.95 are marked with asterisks (*).

τ	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0.0	$-0.017\,692$	$-0.013\,808$	$-0.006\,068$	$-0.008\,696$
0.2	$0.288\,29$	$-0.031\,35$	$-0.039\,795$	$0.148\,92$
0.4	$0.515\,53$	$-0.041\,593$	$-0.010\,872$	$0.222\,59$
0.6	$0.591\,24$	$0.005\,305$	$0.037\,288$	$0.261\,54$
0.8	$0.606\,07$	$0.006\,205$	$0.020\,356$	$0.262\,63$

Table 19: Burden effect: bias in estimating the optimal policy parameter at sample size 500. The algorithm uses λ^* instead of learning λ online. Bias$=\mathbb{E}(\hat{\theta}_t) - \theta^*$.

τ	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0.0	$0.029\,022$	$0.016\,576$	$0.015\,445$	$0.016\,196$
0.2	$0.120\,73$	$0.022\,334$	$0.021\,348$	$0.042\,485$
0.4	$0.294\,46$	$0.018\,117$	$0.015\,525$	$0.065\,667$
0.6	$0.366\,97$	$0.011\,343$	$0.011\,526$	$0.078\,681$
0.8	$0.378\,72$	$0.008\,136$	$0.007\,766$	$0.076\,209$

Table 20: Burden effect: MSE in estimating the optimal policy parameter at sample size 500. The algorithm uses λ^* instead of learning λ online.

τ	θ_0	θ_1	θ_2	θ_3
0.0	0.944	0.950	0.952	0.933*
0.2	0.689*	0.943	0.959	0.815*
0.4	0.159*	0.944	0.954	0.6*
0.6	0.006*	0.941	0.928*	0.295*
0.8	0*	0.94	0.944	0.144*

Table 21: Burden effect: coverage rates of percentile-t bootstrap confidence intervals for the optimal policy parameter at sample size 500. The algorithm uses λ^* instead of learning λ online. Coverage rates significantly lower than 0.95 are marked with asterisks (*).

τ	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	0.392	0.3723	0.3713	−0.0006
0.2	0.3921	0.3722	0.3713	−0.0006
0.4	0.392	0.3723	0.3713	−0.0006
0.6	0.392	0.3723	0.3713	−0.0006
0.8	0.392	0.3723	0.3713	−0.0006

Table 22: Burden effect: the myopic equilibrium policy.

F Nonlinear Reward: The Optimal Policy

α	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	0.418 035	0.395 067	0.397 071	−0.001 615
0.2	0.496 240	0.296 973	0.385 421	0.000 480
0.4	0.564 503	0.202 857	0.365 239	0.001 684
0.6	0.811 000	0.542 000	0.888 000	0.925 000

Table 23: Nonlinear reward: the optimal policy.