An Actor-Critic Contextual Bandit Algorithm for Personalized Interventions using Mobile Devices

Huitian Lei
Department of Statistics
University of Michigan
Ann Arbor, MI 48109
ehlei@umich.edu

Ambuj Tewari
Department of Statistics and Department of EECS
University of Michigan
Ann Arbor, MI 48109
tewaria@umich.edu

Susan Murphy
Department of Statistics, Department of Psychiatry and Institute for Social Research
Ann Arbor, MI 48109
samurphy@umich.edu

Abstract

An Adaptive Intervention (AI) personalizes the type, mode and dose of intervention based on users' ongoing performances and changing needs. A Just-In-Time Adaptive Intervention (JITAI) employs the real-time data collection and communication capabilities that modern mobile devices provide to adapt and deliver interventions in real-time. The lack of methodological guidance in constructing data-based high quality JITAI remains a hurdle in advancing JITAI research despite the increasing popularity JITAIs receive from clinical and behavioral 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. However, interpretability concerns lead us to formulate the problem differently from existing formulations intended for web applications such as ad or news article placement. 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 actor-critic algorithm, including consistency and rate of convergence of reward and JITAI parameters are provided and verified by a numerical experiment. To the best of our knowledge, our is the first application of the actor-critic architecture to contextual bandit problems.

1 Introduction

Advanced technology in mobile devices including smartphones and wearable devices provides a great platform for the delivery of JITAIs (Just in time adaptive intervention, [1]). Adaptive intervention (AI), as an improvement to one-size-fits-all interventions, are interventions designed to adapt to users' heterogeneous nature, preferences, ongoing performances, changing needs, etc [2, 3]. An efficacious intervention for a particular user given a particular context may not be as efficacious for others, and may not work well at a different time point given a different context. JITAI is the real time version of AI. While traditional AI has restriction on the time, location and frequency that an intervention is delivered (e.g., patients' visit to doctor's office), JITAIs allow for more flexibility in intervention delivery. That is, both the adaptation and delivery of intervention happens in real-time via users' mobile devices.

JITAIs have enjoyed wide popularity among behavioral scientists for supporting health behavioral change. An increasing number of studies have been dedicated to assess the effect of JITAI on regulating human health behavior. As an example, a study by Witkiewitz et al. [4] evaluates a mobile feedback intervention targeting heavy episodic drinking and smoking. The focuses of other studies include physical activity [5, 6], obesity/weight management [7], etc. For a comprehensive list of studies, see [8].

Despite the practical appeal and promise of JITAIs, there is a lack of theory and methodology that guides the development and refinement of JITAIs. The construction of a JITAI in the studies listed above is, by and large, solely based on domain behavioral theory. While a theory-based JITAI may serve a good staring point, it is desirable to utilize data collected from the study to improve the theory-based JITAI. In fact the improved JITAI may even help refine and renew the existing domain theories. In this article, we make a first attempt to provide a framework to construct and refine a JITAI. In Section 2 we formulate our problem as an online contextual bandits learning problem and focus a class of parametrized stochastic JITAIs. Section 3 presents an actor-critic algorithm where the critic iteratively learns the average reward and the actor optimizes the average reward in search for a better JITAI. Section 4 establishes the theoretical result on the actor-critic algorithm's consistency and the rate of convergence to the optimal JITAI. Section 5 provides some numerical studies.

2 Problem formulation

We will first pose the problem of learning a JITAI in real-time as an online contextual bandit problem. Next, we argue for the need for a low dimensional parameterization for stochastic policies and motivate the need for further regularization of policy parameters even though they may already be low-dimensional.

2.1 A contextual bandit formulation with linear expected reward

The problem to construct and refine a JITAI can be formulated as an online stochastic multi-armed bandit problem with context or side information, which we refer to as "contextual bandit" [9] throughout this article. When providing JITAIs to users via mobile devices, we consider a series of pre-chosen decision points $\{1, 2, ..., t, ...\}$ where user-device interaction happens.

1. The finite action space $\mathcal{A}$ at any decision point t consists of all suitable interventions. This may include different types of intervention, different dose of intervention, and different modalities of intervention delivery.
2. The contexts s_t at decision point t is a vector of summary information collected by the mobile devices which are relevant for intervention assignment. Depending on the domain of application, context may include GPS location, use calendar data, step counts, local weather, and user self-reported measures. Denote the context space by $\mathcal{S}$.
3. When a context s_t is observed, an algorithm chooses an action a_t and a reward r_t is revealed before the next decision point. The expectation of the reward depends on both the context and the action. We consider the scenario where the expected reward given context s and action a is a linear function of a d dimensional reward feature vector $f(s, a)$ with a unknown coefficient vector μ^*.

$$E(R|s, a) = f(s, a)^T \mu^* \tag{1}$$

   The choice of a_t is based on the algorithm's analysis of context-action-reward tuples from the previous $t-1$ decision points. Upon acquiring the new information at decision point t, the algorithm updates its analysis in order to improve its choice of action at decision point $t+1$.
4. A JITAI, which takes as input possibly all user information available up the decision point and outputs an action or a probability distribution on action space, can be viewed as a policy in reinforcement learning terminology. We will use the term policy in place of JITAI from now on.

2.2 Parameterized stochastic policy

A large body of the existing literature on contextual bandits is devoted to developing online algorithms that achieve good online performances in terms of regret. The action at a decision point t is chosen either by comparing the upper confidence bounds or the posterior distribution for mean reward of different arms (e.g., [10, 11]). Fixing an algorithm, the policy maps the entire trajectory $\{s_i, a_i, r_i\}_{i=1}^{t-1}$ to a particular action.

In this article, we consider a class of parameterized stochastic policies $\pi_\theta(s, a)$ where $\theta \in \mathbb{R}^p$ is a p dimensional vector. Given a fixed parameter θ, the function $\pi_\theta(s, a)$ maps each context s in the context space $\mathcal{S}$ to a probability distribution on the action space $\mathcal{A}$. The reason to consider parameterized stochastic policies come is twofold. First, parameterized policies are interpretable in the sense that the role the context has in determining the probability of choosing different actions is easy to grasp by looking at the policy parameters. Different factors contribute differently to action probabilities depending on the signs and relative magnitude of the corresponding θ_i. Second, the randomness involved in stochastic policies helps creating variety of intervention, which is necessary and effective in preventing intervention burden/habituation/boredom (e.g., [12]). We, in particular, restrict our attention to the class of logistic JITAIs:

$$\pi_\theta(s, a) = \frac{e^{g(s,a)^T\theta}}{1 + e^{g(s,a)^T\theta}} \tag{2}$$

where $g(s, a)$ is a p dimensional vector that jointly summarizes context and action.

2.3 Optimal policy and regularized average reward

The average reward of a policy $\pi_\theta(s, a)$ is the expected reward $E(R|s, a)$ weighted by the policy-specified probability distribution over action space and the distribution over context space.

$$V(\theta) = \sum_{s\in\mathcal{S}} d(s) \sum_{a\in\mathcal{A}} E(R|s, a)\pi_\theta(s, a)$$

Naturally, we define the optimal policy π_{θ^*} as the policy that maximizes the average reward, $\theta^{**} = \operatorname{argmax} V(\theta)$. An online algorithm should therefore be designed to ensure eventual convergence to this optimal policy. By taking a closer inspection, however, we notice that there are two issues when defining the optimal policy to be the maximizer of average reward.

First, consider the scenario where there is zero or negligible intervention effects. This means the expected reward $E(R|s, a)$ depends on the context but is independent of the action. An intuitive and reasonable optimal policy, in this case, is the pure stochastic policy that assigns a uniform probability distribution on the action space $\mathcal{A}$. However, the average reward function $V(\theta)$ now becomes a constant, maximizing which is not only an ill-posed problem [13], but also fails to converge to the pure stochastic policy.

Second, consider a toy example where the action space $\mathcal{A} = \{-1, 1\}$ and context space $\mathcal{S} = \{-1, 1\}$ follows a Bernoulli distribution with $p(s = 1) = p_1$. Assuming the class of logistic policy has the form $\pi_\theta(s, a = 1) = \frac{e^{\theta_0+\theta_1 s}}{1+e^{\theta_0+\theta_1 s}}$ and $\pi_\theta(s, a = 0) = 1 - \pi_\theta(s, a = 1)$, the average reward function becomes

$$V(\theta) = p_1 \frac{r_{1,1}e^{\theta_0+\theta_1} + r_{1,-1}}{1 + e^{\theta_0+\theta_1}} + (1 - p_1)\frac{r_{1,-1}e^{\theta_0-\theta_1} + r_{-1,-1}}{1 + e^{\theta_0-\theta_1}}$$

where $r_{s,a} = E(R|s, a)$. Straightforward calculus leads to the observation that the optimal θ^{**} that maximizes $V(\theta)$ either has $\theta_0^{**} = \pm\infty$ or $\theta_1^{**} = \pm\infty$ as long as $r_{1,1} \neq r_{1,-1}$ and $r_{-1,1} \neq r_{-1,-1}$. Similar examples can be provided when the context space is continuous. The toy example illustrates that in the presence of intervention effect, it is likely that policy parameter that maximizes

the average reward has one or more infinitely large θ_i's, which renders all other θ_i's irrelevant and the maximization problem ill-posed.

Our solution to circumvent ill-posed problems is to regularize the average reward by subtracting a L_2 penalty term. The regularized average reward is

$$J(\theta) = \sum_{s \in \mathcal{S}} d(s) \sum_{a \in \mathcal{A}} E(R|s,a)\pi_\theta(s,a) - \lambda \|\theta\|_2^2 \tag{3}$$

where λ is a tuning parameter that controls the amount of penalization. The optimal policy is $\pi_{\theta^*}(s,a)$ with $\theta^* = \operatorname{argmax} J(\theta)$. Maximizing $J(\theta)$ becomes a well-posed problem and convergence to a pure stochastic policy is guaranteed in the absence of intervention effects. In addition, assuming both reward and contexts are bounded, we essentially limits the search of optimal policy, now being the maximizer of $J(\theta)$, within a compact subset of $\mathbb{R}^p$ (Lemma 1). A minimum exploration probability is therefore guaranteed with $g(s,a)^T\theta$ bounded.

Remark. Note that the use of regularization above is different from situations where regularization is used to prevent overfitting (when learning from finite data) and is only present in the empirical objective being optimized. We, in contrast, introduced a penalty in our *population objective*. This makes sure that the optimal policy, at the population level, is completely exploring when there is no treatment effect. It also prevents the optimal policy from having arbitrarily large parameters.

3 An actor-critic algorithm with linear expected reward

Algorithm 1: Actor critic algorithm for linear expected reward

T_{max} is the total number of decision points.
Critic initialization: $B(0) = I_{d\times d}$, a $d \times d$ identity matrix. $A(0) = 0_d$, $\mu_0 = 0_d$, are $d \times 1$ column vectors.
Actor initialization: θ_0 to be the theory based policy parameter provided by behavioral scientists.
Start from $t = 0$.
while $t < T_{max}$ **do**
At decision point t, observe context s_t ;
Draw an action a_t according to probability distribution $\pi_{\theta_{t-1}}(s_t, a)$;
Observe an immediate reward r_t ;
Critic update:
$B(t) = B(t-1) + f(s_t,a_t)f(s_t,a_t)^T$, $A(t) = A(t-1) + f(s_t,a_t)r_t$, $\mu_t = B(t)^{-1}A(t)$. ;
Actor update:

$$\theta_t = \underset{\theta}{\operatorname{argmax}} \frac{1}{t} \sum_{\tau=1}^{t} \sum_{a} f(s_\tau, a)^T \mu_t \pi_\theta(s_\tau, a) - \lambda \|\theta\|_2^2 \tag{4}$$

Go to decision point $t + 1$.
end

We presents an actor-critic algorithm for learning the optimal policy under the linear expected reward structure. The critic performs a least-squares-like iterative update in estimating the unknown reward parameter μ, with the addition an identity matrix to guarantee matrix $B(t)$ is invertible. The actor uses the critic's estimate μ_t to obtain a plug-in estimate of the expected reward and thus an estimate of the regularized reward function $J(\theta)$. The actor then solves the maximization problem using a global optimization solver such as Baron [14, 15], due to the non-convexity of the objective function. The maximizer is stored as the updated policy parameter.

4 Theory: consistency and rate of convergence

The following two theorems establish the consistency and rate of convergence of both the reward parameter μ_t and the policy parameter θ_t from the actor critic algorithm.

Theorem 1. *The critic's estimate μ_t converges to μ^* in probability. The convergence rate is $O(1/\sqrt{t})$, the optimal parametric convergence rate.*

The main difficulty in proving consistency and rate of convergence of μ_t comes from the non-iid distribution of (s_i, a_i), since a_i is generated using policy $\pi_{\theta_{i-1}}(s_i, a)$ that depends on history $\{s_j, a_j, r_j\}_{j=1}^{i-1}$. Our proof consists of constructing martingale difference sequences and utilizing the matrix version of Azuma's inequality [16]. The sketch of proof in appendix shows more details. Assuming that the global maximizer of the regularized average reward is unique and well-separated from other local maximizers (assumption 3), we have

Theorem 2. *The critic's estimate θ_t converges to θ^* in probability. The convergence rate is $O(1/\sqrt{t})$.*

The gap between θ_t and θ^* comes in two parts. First, θ_t is calculated based on an empirical version of the objective function $J(\theta)$. Second, due to the lack of knowledge of μ^*, θ_t is calculated by replacing μ^* by its estimate μ_t. We prove theorem 2 by bounding both parts of the gap.

We shall point out that, when focusing on a class of parameterized policies, analyzing rate of convergence of policy parameter is closely related to regret analysis in the existing contextual bandit literature where the policy class is unrestricted. The expected accumulated average reward up to decision point t is $\sum_{t=1}^{T}\sum_{a\in\mathcal{A}} f(s_t,a)^T\mu^*\pi_{\theta_{t-1}}(s_t,a)$ for the actor critic algorithm. The expected average reward is $\sum_{t=1}^{T}\sum_{a\in\mathcal{A}} f(s_t,a)^T\mu^*\pi_{\theta^*}(s_t,a)$ had we taken the optimal policy. The regret incurred by the algorithm is

$$\text{Regret}(T) = \sum_{t=1}^{T}\sum_{a\in\mathcal{A}} f(s_t,a)^T\mu^*(\pi_{\theta_{t-1}}(s_t,a) - \pi_{\theta^*}(s_t,a)) \tag{5}$$

which can be bounded by $\sum_{t=1}^{T}\|\theta_t - \theta^*\|_2$ up to a constant. $\text{Regret}(T)$ can thus be bounded by $t^{1-\alpha}$ if θ_t converges to θ^* at the rate of $t^{-\alpha}$. In our case, the regret can be bounded by $\sqrt{T}$ up to a constant.

Corollary 1. *The regret of the online actor-critic algorithm satisfies:*

$$\text{Regret}(T) = \mathcal{O}(\sqrt{T}).$$

5 Numerical experiment

To confirm a convergence theory in the previous section, we test the actor critic algorithm using a hypothetical example in the context of reducing to smoking for heavy smokers. We wish to emphasize that the example we present is a simplification of a real-life scenario and we do not claim any scientific validity of the particular models it uses.

A user is equipped with a smartphone, which prompts the user three times a day for completion of an ecological momentary assessment [17]. Contextual information available from EMA includes user self-reported smoking urge and negative affect. The completion of each EMA questionnaire opens an opportunity to intervene smoking behavior. While EMA itself is not considered as an effective intervention to reduce smoking, we have the option to provide an ecological momentary intervention (EMI, [18]) such as urge-surfing [4]. The two available actions, at any decision point, are "do not provide EMI" (coded as 0) and "provide EMI" (coded as 1). The momentary outcome C, defined to be the smoking rate (per hour) till the next EMA prompt, is generated according to the following linear form given context s and action a

$$C = E(C|S=s, A=a) + \epsilon = \beta_0^T s + a\beta_1^T s + \epsilon$$

where $s = [1, s_1, s_2, s_1 s_2]$ is a 4 dimensional reward feature. s_1 represents smoking urge and s_2 represents negative affect with higher value indicating higher urge and more negative affect. Both s_1 and s_2 are standardized to range between 0 and 1. The noise term ϵ is 0.1 times a truncated standard

normal at ± 1. We choose $\beta_0 = [1.2, 0.8, 0.5, 0]$ and $\beta_1 = [-0.5, -0.2, +0.3, -0.15]$. The choice of β_0 reflects that in general both higher smoking urge and negative affect are associated with increased cigarette consumption. The choice of β_1 reflects a theory that EMI helps reducing smoking on average with increased intervention effect, the reduction in smoking when EMI is provided compared to no EMI, when smoking urge is high and diminished effect when negative affect is high. Given a particular value of negative affect, the intervention effect increases with the value of smoking urge. This outcome, smoking rate, is considered as negative reward, or cost, in the context of smoking reduction, thus the policies and algorithms are designed to minimize the regularized average cost.

$$\tilde{J}(\theta) = \sum_{s \in \mathcal{S}} d(s) \sum_{a \in \mathcal{A}} E(C|s,a)\pi_\theta(s,a) + \lambda \|\theta\|_2^2$$

In this simplified experiment set up, the policy feature coincides with the reward feature. The total number of decision points is 300, which corresponds to 3 EMAs per day for about 3 consecutive months. We choose $\lambda = 0.1$. We find that *fminunc* in MATLAB provides reliable global maximum in our experiment, so the results below are generated using *fminunc*. Figures represent one trajectory of experiment since others behavior similarly. Figure 1 and Figure 2 verifies the consistency of μ_t and θ_t. We emphasis that while μ_t does not converge well for 300 decision point ($\|\mu_{300} - \mu^*\|_2 = 0.1387$) compared to the convergence if extended experiment was run ($\|\mu_{1000} - \mu^*\|_2 = 0.0321$), the policy parameter converges reasonably well. Figure 3 shows that the regularized average cost $\tilde{J}(\theta_t)$ of online policies converge to that of the optimal policy. Figure 4 verifies the $\sqrt{t}$ convergence rate by showing that the time-scaled square L_2 distance $t\|\theta_t - \theta^*\|_2^2$ is bounded.

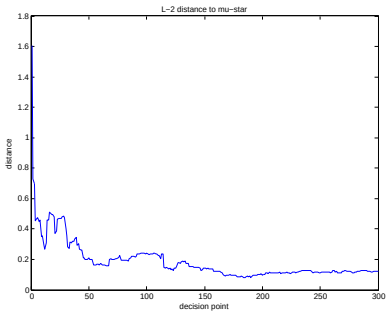

Figure 1: L_2 distance between μ_t and true reward parameter μ^*

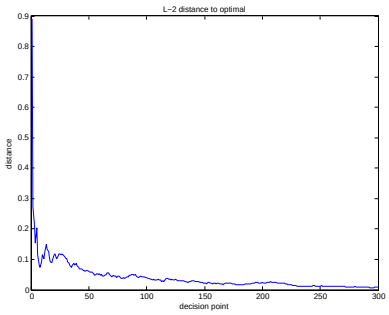

Figure 2: L_2 distance between θ_t and optimal policy parameter θ^*

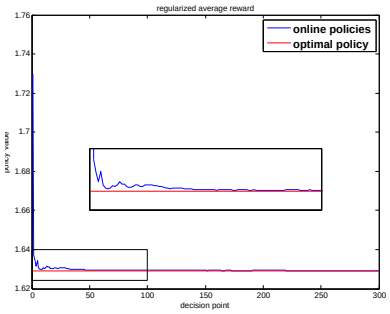

Figure 3: regularized average cost of online policies

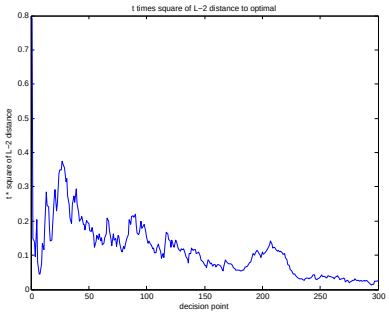

Figure 4: time-scaled square L_2 distance $t\|\theta_t - \theta^*\|_2^2$

Acknowledgments

We acknowledge the support of University of Michigan Mcubed grant for Methodology for Developing Real-Time Mobile Phone Adaptive Interventions, NIH grants P50DA010075, R01MH099898, R01HD073975.

6 Appendix

Before proceeding to proof of Theorem 1 and Theorem 2, we introduce the follow functions that will be used in the proof. $g(\theta, \mu)$ presents the regularized average reward of policy π_θ when the reward parameter is μ, a mapping from the product space of policy parameter and reward parameter, $\mathbb{R}^{p+d}$ to $\mathbb{R}$.

$$g(\theta, \mu) = \sum_s d(s) \sum_a f(s,a)^T \mu \pi_\theta(s,a) - \lambda \|\theta\|_2^2$$

The random function $G(\theta, \mu, S)$ where context S is the random variable, measures the regularized average reward under policy π_θ and reward parameter μ when observing context S.

$$G(\theta, \mu, S) = \sum_a f(S,a)^T \mu \pi_\theta(S,a) - \lambda \|\theta\|_2^2$$

We denote the empirical average of the random functions by

$$\bar{G}_t(\theta, \mu) = \frac{1}{t} \sum_{\tau=1}^{t} \sum_a f(S_\tau, a)^T \mu \pi_\theta(S_\tau, a) - \lambda \|\theta\|_2^2$$

Assumption 1. *Boundedness: Both the reward feature and the policy feature are bounded. In particular $\|f(s,a)\|_2 \leq 1$ and $\|g(s,a)\|_2 \leq 1$ with probability one. The rewards R are bounded with probability one.*

Assumption 2. *The $d \times d$ matrix $\sum_s d(s) \sum_a f(s,a) f(s,a)^T$ is positive definite.*

Lemma 1. *There exist $K > 0$ such that $\|\theta^*\|_2 < K$, and $\|\theta_t\|_2 < K$ with probability 1, which means that the maximization problem (4) is essentially restricted to a compact set in $\mathbb{R}^p$.*

Proof. The proof is straightforward by noticing that for θ^* to be the maximizer of $g(\theta, \mu)$ when $\mu = \mu^*$, it is necessary that $g(\theta^*, \mu^*) \geq g(0, \mu^*)$. □

We first prove that critic's estimate μ_t converges to μ^* in probability.

Proof. To prove that $|\mu_t - \mu^*|^2 = \tilde{A}(t)\tilde{B}(t)^{-1}\tilde{B}(t)^{-1}\tilde{A}(t)$ converge to 0 in probability, it is sufficient to prove that (1) $\frac{\tilde{B}(t)^{-1}}{t}$ is bounded with probability going to one and (2)$\frac{\tilde{A}(t)}{t}$ converges to 0 in probability. Here $\tilde{B}(t) = I_d + \sum_{i=1}^t f(s_i, a_i) f(s_i, a_i)^T$, $\tilde{A}(t) = \sum_{i=1}^t f(s_i, a_i)\epsilon_i - \mu^*$ where $\epsilon_i = r_i - f(s_i, a_i)^T \mu^*$ is the noise term. Both claim (1) and (2) are proved by constructing suitable martingale difference sequences and using Azuma's and matrix Azuma's inequality [16]. □

We make two additional assumptions in proving the first part of Theorem 2. The first assumption concerns the seperation of the global maxima of $g(\theta, \mu)$ from other local maximas. The second assumptions concerns the smoothness of function $g(\theta, \mu)$. We should point out that having assumed that both the reward and policy features are bounded, assumption 4 is actually a direct consequence of the boundedness.

Assumption 3. *The global maximum is uniformly well separated from other local maximum. For any fixed μ, use the set S^μ to denote the collection of θ that produces a local maximum of $g(\theta, \mu)$.*

The global maximizer θ^μ, of course, belongs to S^μ. We assume that there exists a constant δ, free of μ, such that

$$\max_{\theta \in S^\mu} g(\theta, \mu) - \max_{\theta \in S^\mu, \theta \neq \theta^\mu} g(\theta, \mu) \geq \delta$$

for all μ.

Assumption 4. *The Jacobian matrix $g''_{\theta\theta}$ is uniformly Lipschitz in θ. There exists a constant vector L, free of μ, such that*

$$\lambda_{max}(g''_{\theta\theta}(\theta_1, \mu) - g''_{\theta\theta}(\theta_2, \mu)) \leq L^T \|\theta_1 - \theta_2\|_2$$
$$\lambda_{min}(g''_{\theta\theta}(\theta_1, \mu) - g''_{\theta\theta}(\theta_2, \mu)) \leq L^T \|\theta_1 - \theta_2\|_2$$

for all μ. Or equivalently, we assume that $g'''_{\theta\theta\theta}$ is universally bounded.

Next we prove that actor's estimate θ_t converges to θ^* in probability.

Proof. The proof on consistency of θ_t leverages on the consistency of μ_t and comes in two step. We introduce an intermediate variable $\tilde{\theta}_t = argmax_\theta g(\theta, \mu_t)$. We first show that given the consistency of μ_t, $\tilde{\theta}_t$ converges to θ^* by applying Lemma 9.1 in [19]. The second step is to prove that the M estimator converges uniformly in a neighborhood of μ^*, namely

$$\theta_t^\mu = argmax_\theta \bar{G}_t(\theta, \mu) \to \theta^\mu = argmax_\theta g(\theta, \mu)$$

in probability, and uniformly over all μ in a neighborhood of μ^*. The proof follows essentially the same paradigm as Theorem 9.4 in [19] and is omitted here. □

Next we prove that $tE(\mu_t - \mu^*)(\mu_t - \mu^*)^T$ converges to a finite matrix, which completes the proof of Theorem 1. The proof again hinges on constructing proper martingale difference sequences and utilizing matrix Azuma's inequality and therefore is omitted here.

Last but not least, we complete the proof of Theorem 2 by showing that θ_t converges to θ^* in a $O(1/\sqrt{t})$ rate.

Proof. The proof of rate of convergence of θ_t again leverages on that of μ_t. We first show that $\tilde{\theta}_t$ converges to μ^* in square root rate. The expression

$$\theta^* - \tilde{\theta}_t = (\tilde{g}''_{\theta\theta})^{-1} \tilde{g}''_{\theta\mu}(\mu^* - \mu_t)$$

is obtained by Taylor expanding $g'_\theta(\theta^*, \mu^*) = 0$ and $g'_\theta(\tilde{\theta}_t, \mu_t) = 0$. $\tilde{g}''_{\theta\theta}$ and $\tilde{g}''_{\theta\mu}$ are partial derivatives evaluated at some intermediate point between (θ^*, μ^*) and $(\tilde{\theta}_t, \mu_t)$. The $O(1/\sqrt{t})$ convergence rate then follows from assumption 4.

The second step is to show that θ_t^μ converges to θ^μ, uniformly over μ in a neighborhood of μ^*. Using the expression

$$0 = \sqrt{t} \int G'_\theta(\theta, \mu, S) d(\mathbb{P}_t - \mathbb{P}) + \sqrt{t}[g''_{\theta\theta}(\theta^\mu, \mu) + g'''_{\theta\theta\theta}(\theta_t^\mu - \theta^\mu)\alpha_\mu](\theta_t^\mu - \theta^\mu)$$

and the fact that the class of function $G'_\theta(\theta, \mu, S)$ is P-Donsker [20], the uniform convergence follows though.

□

References

[1] William T Riley, Daniel E Rivera, Audie A Atienza, Wendy Nilsen, Susannah M Allison, and Robin Mermelstein. Health behavior models in the age of mobile interventions: are our theories up to the task? *Translational behavioral medicine*, 1(1):53–71, 2011.

[2] Linda M Collins, Susan A Murphy, and Karen L Bierman. A conceptual framework for adaptive preventive interventions. *Prevention science*, 5(3):185–196, 2004.

[3] Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

[4] Katie Witkiewitz, Sruti A Desai, Sarah Bowen, Barbara C Leigh, Megan Kirouac, and Mary E Larimer. Development and evaluation of a mobile intervention for heavy drinking and smoking among college students. *Psychology of Addictive Behaviors*, 28(3):639, 2014.

[5] Abby C King, Eric B Hekler, Lauren A Grieco, Sandra J Winter, Jylana L Sheats, Matthew P Buman, Banny Banerjee, Thomas N Robinson, and Jesse Cirimele. Harnessing different motivational frames via mobile phones to promote daily physical activity and reduce sedentary behavior in aging adults. *PloS one*, 8(4):e62613, 2013.

[6] Sunny Consolvo, David W McDonald, Tammy Toscos, Mike Y Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, et al. Activity sensing in the wild: a field trial of ubifit garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1797–1806. ACM, 2008.

[7] Kevin Patrick, Fred Raab, Marc A Adams, Lindsay Dillon, Marian Zabinski, Cheryl L Rock, William G Griswold, and Gregory J Norman. A text message–based intervention for weight loss: randomized controlled trial. *Journal of medical Internet research*, 11(1), 2009.

[8] Susan A.Murphy Inbal Nahum-Shani, Shawna N.Smith. Just in time adaptive interventions (jitais): An organizing framework for ongoing health behavior support. submitted.

[9] John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824, 2008.

[10] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

[11] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. *arXiv preprint arXiv:1209.3352*, 2012.

[12] Ben S Gerber, Melinda R Stolley, Allison L Thompson, Lisa K Sharp, and Marian L Fitzgibbon. Mobile phone text messaging to promote healthy behaviors and weight loss maintenance: a feasibility study. *Health informatics journal*, 15(1):17–25, 2009.

[13] Michael W Mahoney. Approximate computation and implicit regularization for very large-scale data analysis. In *Proceedings of the 31st symposium on Principles of Database Systems*, pages 143–154. ACM, 2012.

[14] M. Tawarmalani and N. V. Sahinidis. A polyhedral branch-and-cut approach to global optimization. *Mathematical Programming*, 103:225–249, 2005.

[15] N. V. Sahinidis. *BARON 12.1.0: Global Optimization of Mixed-Integer Nonlinear Programs,* User's Manual, 2013.

[16] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[17] Saul Shiffman and Arthur A Stone. Ecological momentary assessment: A new tool for behavioral medicine research. *Technology and methods in behavioral medicine*, pages 117–131, 1998.

[18] Kristin E Heron and Joshua M Smyth. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. *British journal of health psychology*, 15(1):1–39, 2010.

[19] Robert W. Keener. *Theoretical statistics: Topics for a Core Course*. New York, NY : Springer Science+Business Media, 2010.

[20] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.