Robustness of the Contextual Bandit Algorithm to A Physical Activity Motivation Effect

Xige Zhang

April 10, 2016

1 Introduction

Technological advances in mobile devices have seen a growing popularity in Just-In-Time Adaptive Interventions (JITAI), which are interventions that are adapted and delivered in real-time to reflect individuals' behaviors and needs in their daily lives [1]. Compared to traditional Adaptive Interventions, JITAI is able to adapt and deliver interventions in real-time according to present contextual variables by taking advantage of the information observed by individuals' mobile devices. Many mobile applications, such as personalized news article recommendations, have successfully applied JITAI to adapt their recommendations to people's interests according to the time value of news in real-time [2]. Given the power of JITAI, we would like to apply it to mobile health problems.

Previous studies have shown that JITAIs can support behavior changes in many fields, such as eating disorder [3] and smoking cessation [4]. An example of JITAI in mobile health is HeartSteps [5], which is a mobile intervention application installed on Android smartphones and

designed to help cardiac patients maintain physical activity after their cardiac rehabilitation [6]. HeartSteps adapts personalized suggestions for individuals' physical activity through frequent interactions with individuals via their phones. Suggestions are delivered in real-time with intervention policies, given patients' real-time contextual variables, such as weather, time, and location. Provided the features of HeartStep, it is suitable to model the JITAI as a contextual bandit problem for the purpose of learning the optimal treatment plan and delivering immediate interventions on mobile devices.

A contextual bandit problem is a sequential decision-making problem [1]. At each decision point, its decision making strategy will update simultaneously according to a learning algorithm that observes side information, interventions and feedbacks. The value of a new context is evaluated by exploring a feedback collected (exploration), and an optimized decision is provided in real-time given the current information (exploitation). A trade-off between exploration and exploitation would reveal an optimal policy for decision making. In mobile health problems, we want to take advantage of the contextual bandit algorithm to provide optimal physical activity suggestions for people.

However, in the real world application of learning the optimal policy for recommending physical activities to people, deviations to the algorithms' assumptions occur. It is traditionally assumed that the distribution of context states at a decision point is independent of the interventions, contexts and costs from previous decision points [1]. In real world mobile health cases, this assumption is fragile, for there are many ways the distribution of contexts can be impacted by past interventions. In general, a physical activity motivation effect is observed as a main factor affecting the distribution in mobile health problems. For example, people may cultivate healthier habits after receiving a recommendation. In order to test the applicability of the contextual bandit algorithm in mobile health problems, this paper investigates the performance of the algorithm when a physical activity motivation effect is present. In particular, it explores the performance of the algorithm when previous interventions impact the distribution of contexts.

Physical activity motivation effects are involved in the context generating procedures. This study focuses on two conditions: when the assumption of the algorithm is satisfied, and when the physical activity motivation effect breaks the algorithm's assumption. We want to compare the performance of the algorithm under these two conditions.

In this paper, we will overview the study procedure in section 2. Section 3 provides an online Actor-Critic algorithm of our learning algorithm [1], which is used for choosing an optimized decision during the process. In section 4, we design the simulation models to the problem under different conditions. Section 5 analyzes and reports the experiment results. Section 7 provides a discussion of the application of the contextual bandit algorithm in mobile health problems.

2 Problem Overview

2.1 The contextual bandit problem

There are four vital elements for a contextual bandit problem [7]:

a. The first element is a series of decision points, {1,2, . . . t, . . . }, when interventions are chosen.

b. The second element is contextual information, based on which interventions are chosen.

c. The third element is a list of sequential intervention decisions.

d. The last element is a decision rule (or policy) for choosing an intervention based on the context at each decision point.

In this study, the problem has a context space $\mathcal{S}$ and a binary intervention space $\mathcal{A} = \{0,1\}$. At each decision point t, context, intervention, and cost are S_t, A_t, and C_t respectively. This study uses a parametrized logistic policy $\pi_\theta(a|s)$ for choosing interventions, given context s (e.g., [5,

8]). To provide a practical meaning of the problem, we interpret the average cost as sedentary time, which we want to minimize. $A_t = 1$ means a physical activity suggestion is provided at time t, while $A_t = 0$ means no suggestion is provided. Contexts are information provided by mobile devices, such as weather and location. Detailed interpretations of context states will be discussed later in section 4. An average cost is observed before the next decision point.

The contextual bandit algorithm is used to find the optimal policy that minimizes the average cost. The purpose of this study is to observe the impact of a physical motivation effect to the performance of the algorithm. In order to limit the search of the optimal policy, and ensure the convergence to the optimal policy, we regularize the average cost by adding a penalty term. The regularized average cost of a policy $\pi_\theta(a, s) = \frac{e^{g(s)^T\theta}}{1+e^{g(s)^T\theta}}$ is a long-term cumulative cost $\mathbb{E}(C|S = s, A = a)$ weighted on the policy-specified probability over context and intervention space [9]. In addition, it subject to a quadratic constraint:

$$J(\theta) = \sum_{s\in\mathcal{S}} d(s) \sum_{a\in\mathcal{A}} \mathbb{E}(C|S = s, A = a)\pi_\theta(a|s) + \lambda\theta^T\mathbb{E}[g(S)^T g(S)]\theta \tag{1}$$

where λ is a tuning parameter that controls the amount of penalization [1]. The problem is to explore an optimal policy (θ^*) such that J(θ) is minimized.

3 A Bandit Actor-Critic Algorithm

The bandit actor-critic algorithm is an online learning of the optimal policy, denoted as θ^*. Starting from t = 0, at each decision point t before T_{max} (the total number of decision points), the Critic algorithm learns about the cost function, and evaluates the quality of a given policy $\pi_\theta(a, s)$ by estimating the average cost together with derivatives. The Actor algorithm uses the information provided by Critic to find the regularized optimal policy. The optimal policy θ^* is the one that minimizes the regularized average cost function J(θ). Both the Critic and the

Actor algorithms update their parameters at each decision point. The strategy of learning the optimal policy is updated at each decision point.

Algorithm 1: An online linear actor critic algorithm

T_{max} is the total number of decision points.

Critic initialization: $B(0) = \zeta I_{d\times d}$, a $d \times d$ identity matrix. $\zeta = 0.001$. $A(0) = 0_d$ is a $d \times 1$ column vector.

Actor initialization: θ_0 is the optimal policy based on domain theory of historical data.

Start from $t = 0$.

while $t \leq T_{max}$ **do**

At decision point t, observe context s_t ;

Draw an intervention a_t according to the probability distribution $\pi_{\theta_{t-1}}(A|S_t)$;

Observe an immediate cost c_t ;

Critic update:

$B(t) = B(t-1) + f(s_t, a_t)f(s_t, a_t)^T$, $A(t) = A(t-1) + f(s_t, a_t)R_t$, $\hat{\mu}_t = B(t)^{-1}A(t)$. ;

Actor update:

$$\hat{\theta}_t = \underset{\theta}{\operatorname{argmin}} \frac{1}{t}\sum_{\tau=1}^{t}\sum_{a} f(S_\tau, a)^T \mu_t \pi_\theta(A = a|S_\tau) + \lambda\theta^T[\frac{1}{t}\sum_{\tau=1}^{t} g(S_\tau, a)^T g(S_\tau, a)]\theta \quad (2)$$

Go to decision point $t + 1$.

end

In the algorithm above, $f(s_t, a_t)$ is the cost feature; in other words, it is the covariates that affect costs for each arm introduced. A small $\zeta = 0.001$ is chosen to ensure a full rank of the initial matrix of B(t), so that it is invertible in the critic update process. The choices for the tuning parameter λ for each model will be discussed in section 4.

4 Simulation Design

The purpose of this study is to explore the performance of the algorithm when context states are non-iid, but depend on previous contexts and/or previous interventions. In this section, we introduce dependency of context states on previous context states and/or previous interventions in different models. By observing the policy values under different models, we can compare the performance of the algorithm under different conditions.

4.1 Generative Model

For each model, we simulate 1000 people, each with a total number of decision points T_{max} = 200 (if not stated specially). The initial distribution of the context states is multivariate normal with mean 0 and covariance matrix

$$\Sigma_1 = \begin{pmatrix} 1 & 0.3 & -0.3 \\ 0.3 & 1 & -0.3 \\ -0.3 & -0.3 & 1 \end{pmatrix}$$

The generative model for context states generation is as follows:

$$\begin{aligned} \xi_t &\sim \text{Normal}_4(0, I), \\ S_{t,1} &= \beta_1 S_{t-1,1} + \xi_{t,1}, \\ S_{t,2} &= \beta_2 S_{t-1,2} + \alpha_2 A_{t-1} x_{coefficient} + \xi_{t,2}, \\ S_{t,3} &= \beta_3 S_{t-1,3} + \gamma_3 S_{t-1,3} A_{t-1} x_{coefficient} + \alpha_3 A_{t-1} x_{coefficient} + \xi_{t,3}, \end{aligned}$$

The context states have different interpretations. $S_{t,1}$ represents weather at decision point t.

$S_{t,2}$ represents the current level of physical activity motivation at decision point t, where the intervention is an engagement to $S_{t,2}$ in particular. The impact of the motivation effect is buried in the contexts simulation process, and the parameter $x_{coefficient}$ measures the size of the motivation effect. $S_{t,3}$ represents a disengagement to an intervention. In order to focus on the impact of $S_{t,2}$, we ignore the effect of $S_{t,3}$ by setting the corresponding coefficient in the cost generation model equal to zero.

In all models below, context states form a vector of length three: $S_t = [S_{t,1}, S_{t,2}, S_{t,3}]$ at each decision point. By setting the coefficients $\beta_1, \beta_2, \beta_3, \alpha_2, \alpha_3, \gamma_3$ to different values, we can observe the performance of the algorithm under different conditions. After the context states are observed, an intervention is chosen according to the parametrized logistic policy:

$$\pi_\theta(A_t = 1 | S_t = [S_{t,1}, S_{t,2}, S_{t,3}]) = \frac{e^{\theta_0 + \theta_1 S_{,t1} + \theta_2 S_{t,2} + \theta_3 S_{t,3}}}{1 + e^{\theta_0 + \theta_1 S_{t,1} + \theta_2 S_{t,2} + \theta_3 S_{t,3}}}$$

Intervention A_t is a binary variable with value 1 or 0. $A_t = 1$ means a physical activity suggestion is provided at time t, while $A_t = 0$ indicates no suggestion at time t. C_t is interpreted as the sedentary time and is revealed according to a linear cost generation model:

$$C_t \;=\; 10 - S_{t,1} - A_t \times (0.25 + 0.25 S_{t,1} + 0.4 S_{t,2}) + \xi_{t,4}$$

where the noise term $\xi_{t,4}$ has i.i.d. standard normal distribution.

The problem of searching for the optimal policy is turned into an unconstrained optimization problem using Lagrangian:

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, J_\lambda(\theta) + \lambda \theta^T \mathbb{E}[g(S)^T g(S)]\theta \tag{3}$$

The stringency of the quadratic constraint increases monotonically as the Lagrangian parameter λ increases [1]. We use line search on a range of λ's to find the smallest one, denoted as λ^*, such

that the minimizer θ^* to the regularized average cost function satisfies the quadratic constraint. We use the same corresponding λ^* in the actor critic learning algorithm.

The non-convexity of the regularized average cost function makes it difficult to solve the optimization problem. Thus, we refine grid search by pattern search method in MATLAB to find the optimal policy for a given λ. The optimal policy value (minimum average cost) for each model is found from the result of 100,000 Monte Carlo samples.

4.1.1 Model.a

In model.a, context states are identically and independently distributed (i.i.d). The assumption of the contextual bandit algorithm is satisfied. Correspondingly, all coefficients in the generative model for contexts simulation are set to zero, which means setting $\beta_1, \beta_2, \beta_3, \alpha_2, \alpha_3, \gamma_3$ all to zeros. The context states are generated as follows:

$$\begin{aligned} \xi_t &\sim \text{Normal}_4(0, I), \\ S_{t,1} &= \xi_{t,1}, \\ S_{t,2} &= \xi_{t,2}, \\ S_{t,3} &= \xi_{t,3}, \end{aligned}$$

When its assumption is satisfied, we expect the contextual bandit algorithm to effectively select the optimal policy. The policy value (average cost) for each simulated person is expected not to deviate far from the optimal policy value.

4.1.2 Model.b

Model.b violates the traditional assumption of the contextual bandit algorithm. We suspect that a physical activity motivation effect would impair the performance of the contextual bandit algorithm. In the models in this section, we add motivation effects in the contexts evolving process. By setting the coefficients in the generative model $[\beta_1, \beta_2, \beta_3, \alpha_2, \alpha_3, \gamma_3]$ to [0.4,0.25, 0.5, 0.8, 0.5, 0.05], we get the following models with different levels of physical activity motivation effects, the size of which is measured by $x_{coefficient}$.

$$
\begin{aligned}
\xi_t &\sim \text{Normal}_4(0, I), \\
S_{t,1} &= 0.4S_{t-1,1} + \xi_{t,1}, \\
S_{t,2} &= 0.25S_{t-1,2} + 0.8A_{t-1}x_{coefficient} + \xi_{t,2}, \\
S_{t,3} &= 0.5S_{t-1,3} + 0.05S_{t-1,3}A_{t-1}x_{coefficient} + 0.5A_{t-1}x_{coefficient} + \xi_{t,3},
\end{aligned}
$$

As a measure of the motivation effect, $x_{coefficient}$ increases from 0 to 1 in the step of 0.2. The impact of motivation effect increases as $x_{coefficient}$ increases.

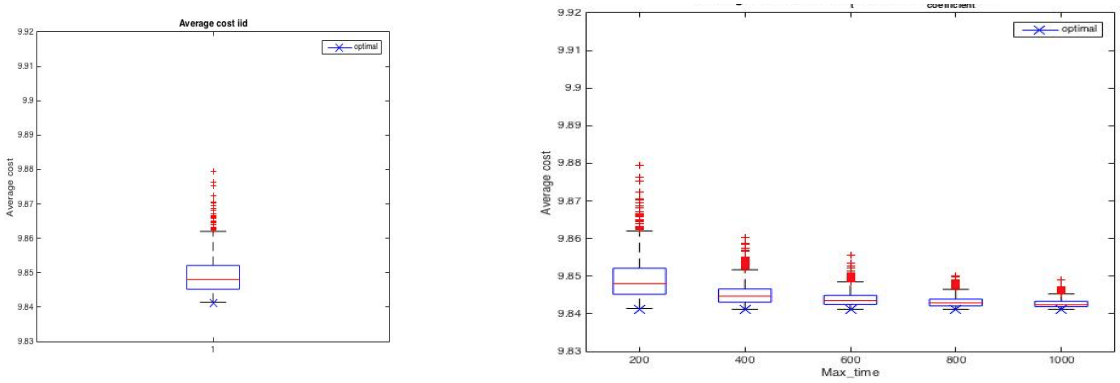
When $x_{coefficient} = 0$, context-simulation is an auto-regressive process. Even though the context states do not depend on the interventions from previous decision points in this case, it still violates the classic bandit problem's assumption.

Intuitively speaking, when a suggestion/intervention is received ($A_{t-1} = 1$), a person gets motivated to do physical activity from the suggestion/intervention. The sedentary time C_t of the person is supposed to decrease. We expect to see a decreasing trend in the average sedentary time as a result of an increasing motivation effect.

5 Simulation Results

5.1 Simulation Results: Model.a

In figure 1, each boxplot is based on 1000 simulated people. Plot (a) shows the simulation results with 200 decision points ($T_{max} = 200$), when the context states are i.i.d. Plot (b) shows the results with different total numbers of decision points T_{max}. Both plot (a) and (b) show that the optimal policy value touches the bottom of each boxplot, and the policy value of each simulated person does not deviate away from the optimal policy value. As the total number of decision points T_{max} increases from 200 to 1000, we observe the variance of the policy values decreasing, and the policy value for each simulated person gets closer to the optimal policy value. This observation suggests that increasing the observation time T_{max} can improve the performance of the contextual bandit algorithm when its assumption is satisfied.



(a) IID when T_{max} is 200

(b) IID as T_{max} changes

Figure 1: IID

Table 1 below provides $\hat{\theta}$'s bias and mean square error (MSE) as T_{max} increases from 200 to 1000, when context states are i.i.d. Bias of $\hat{\theta}$ is calculated by taking the average value of $\hat{\theta}$ over

1000 simulated people, and then subtracting the optimal θ. MSE for each $\hat{\theta}$ is calculated by adding the variance of $\hat{\theta}$ over 1000 simulated people to the square of $\hat{\theta}$'s bias. We observe that both bias and MSE of $\hat{\theta}$ are close to 0, and the MSE values decrease as T_{max} increases.

Table 1: Bias and MSE of Policy Parameters Under i.i.d Model

$\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$

T_{max}	θ_0	θ_1	θ_2	θ_3
200	0.0000	0.0015	0.0055	0.0000
400	-0.0031	0.0014	0.0042	0.0019
600	0.0007	0.0010	0.0037	0.0010
800	0.0015	-0.0004	0.0026	0.0027
1000	0.0001	-0.0009	0.0031	0.0020

$MSE = Var(\hat{\theta}) + Bias(\hat{\theta})^2$

T_{max}	θ_0	θ_1	θ_2	θ_3
200	0.0161	0.0134	0.0127	0.0141
400	0.0082	0.0065	0.0066	0.0068
600	0.0052	0.0045	0.0042	0.0049
800	0.0038	0.0032	0.0031	0.0038
1000	0.0029	0.0026	0.0024	0.0029

5.2 Simulation Result: Model.b

5.2.1 Comparison Between Model.a and Model.b when $x_{coefficient}$ is 0

Figure 2 compares the simulation results of model.a and an auto-regressive model. According to the two boxplots, the policy values are slightly different, and the variances of policy values are similar under the two models. Moreover, table 2 provides the Interquartile Range (IQR) and (Median Absolute Deviation) MAD of policy values for both models, which are also similar. This comparison reveals that the contextual bandit algorithm is still effective in selecting the optimal policy when the contexts generation is auto-regressive.

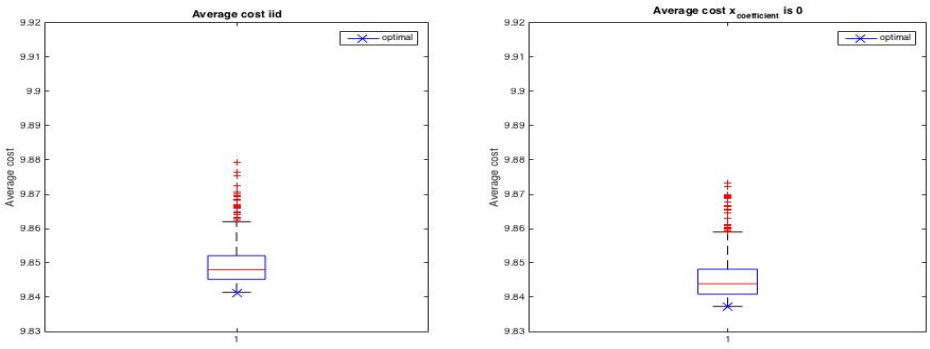


Figure 2: Compare iid and $x_{coefficient}$ is 0

When context states generation is auto-regressive, the policy values are slightly different from when context states are i.i.d. The MAD and IQR are similar under two conditions. According to these observations, even though the auto-regressive context states generation violates the traditional assumption of the contextual bandit algorithm, the performance of the contextual bandit algorithm is still robust.

Table 2: IQR and MDA with Different $x_{coefficient}$

$$IQR = Q_3 - Q_1$$

$$MAD = median_i(|X_i - median_j(X_j)|)$$

Model	i.i.d	Auto-Regressive
IQR	0.0069	0.0073
MAD	0.0043	0.0045

5.2.2 Comparison of Policy Values with Different $x_{coefficient}$

Table 3 provides λ values, optimal policy values, and optimal policy parameters with different levels of physical activity motivation effects measured by $x_{coefficient}$. Larger λ is needed when

$x_{coefficient}$ increases, so that the quadratic constraint is satisfied. We see that θ_0 increases relative to θ_1 and θ_2. Intuitively, it suggests that more motivated people are more likely to exercise regardless of the context. A decreasing trend in the policy value is also revealed in the table when the impact of the motivation effect increases.

Table 3: Optimal policy values and parameters for different $x_{coefficient}$

$x_{coefficient}$	λ^*	optimal policy value	θ_0^*	θ_1^*	θ_2^*	θ_3^*
0	0.13	9.8372	0.2117	0.2078	0.3313	0
0.2	0.16	9.8165	0.2117	0.1688	0.2820	0
0.4	0.18	9.7925	0.2117	0.1531	0.2469	0
0.6	0.21	9.7694	0.2156	0.1352	0.2117	0
0.8	0.22	9.7422	0.2469	0.1313	0.2000	0
1	0.24	9.7161	0.2469	0.1156	0.1844	0

Model.b visualizes the impacts of different levels of physical activity motivation effects on the average sedentary time (average cost). As $x_{coefficient}$ increases, the level of the physical activity motivation effect $S_{t,2}$ increases. According to the cost generation function, an increase in $S_{t,2}$ will decrease the cost C_t. In figure 3, each boxplot is a result of an independent model with a unique $x_{coefficient}$, based on 1000 simulated people with $T_{max} = 200$. The decreasing trend in the average cost values implies that the contextual bandit algorithm captures the impact of the motivation effect. In addition, within each boxplot, the optimal policy value touches the bottom of the corresponding boxplot. These observations indicate that, in our generative model, the contextual bandit algorithm is still robust when a physical activity motivation effect is present.

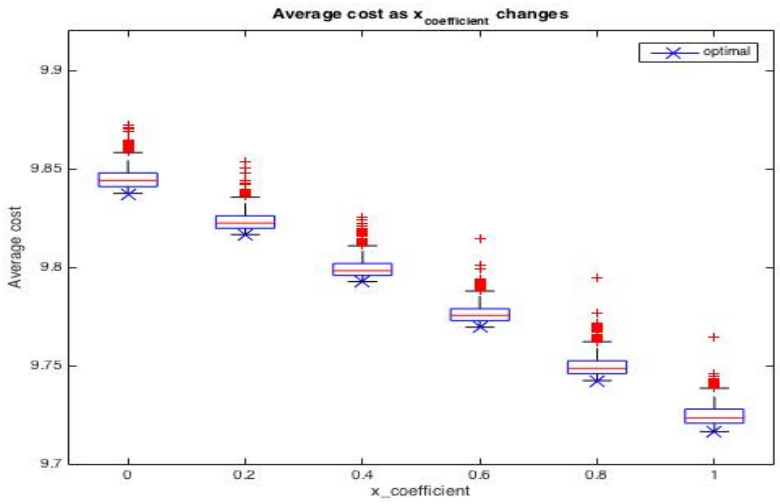


Figure 3: Average Sedentary Time with Different Levels of Physical Activity Motivation

Table 4 indicates that the IQR and MAD of the average cost do not have a significant increase as $x_{coefficient}$ increases. Also, we observe from table 5 that the bias and MSE of the policy values are relatively small and do not increase significantly as $x_{coefficient}$ increases.

Table 4: IQR and MDA with Different $x_{coefficient}$

$$IQR = Q_3 - Q_1$$

$$MAD = median_i(|X_i - median_j(X_j)|)$$

$x_{coefficient}$	0	0.2	0.4	0.6	0.8	1
IQR	0.0073	0.0051	0.0061	0.0056	0.0065	0.0067
MAD	0.0034	0.0025	0.0028	0.0027	0.0029	0.0032

Table 5: Mean Square Error of Average Cost

$$Bias(\hat{C}) = \mathbb{E}(\hat{C}) - C^*$$

$$MSE = Var(\hat{C}) + Bias(\hat{C})^2$$

$x_{coefficient}$	0	0.2	0.4	0.6	0.8	1
Bias	0.0081	0.0062	0.0070	0.0069	0.0077	0.0084
MSE	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001

5.2.3 Policy Parameters With Different $x_{coefficient}$

We tracked the policy parameters θ for different $x_{coefficient}$, and calculated the bias and MSE of $\hat{\theta}$'s. Table 6 below shows that both the bias and MSE of $\hat{\theta}$'s are small, and do not have significant increase as $x_{coefficient}$ increases.

The observations of the policy values and policy parameters for different $x_{coefficient}$ values suggest that the contextual bandit algorithm can still effectively select the optimal policy in our generative model. According to our generative model, the contextual bandit algorithm is still robust when different levels of physical activity motivation effects are present.

Table 6: Bias and MSE of Policy Parameters With Different $x_{coefficient}$

$Bias(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$

$x_{coefficient}$	θ_0	θ_1	θ_2	θ_3
0	-0.0028	0.0023	-0.0116	0.0057
0.2	-0.0313	0.0016	-0.0056	0.0026
0.4	-0.0505	0.0052	0.0048	0.0021
0.6	-0.0739	0.0030	0.0013	0.0017
0.8	-0.1106	0.0003	0.0013	0.0013
1	-0.1193	0.0048	0.0017	0.0015

$MSE = Var(\hat{\theta}) + Bias(\hat{\theta})^2$

$x_{coefficient}$	θ_0	θ_1	θ_2	θ_3
0	0.0154	0.0111	0.0109	0.0112
0.2	0.0091	0.0060	0.0061	0.0057
0.4	0.0119	0.0067	0.0065	0.0062
0.6	0.0133	0.0050	0.0049	0.0047
0.8	0.0202	0.0047	0.0043	0.0041
1	0.0217	0.0042	0.0033	0.0034

5.2.4 Collinearity Issue

In figure 3, we observe significant outliers in boxplots when $x_{coefficient}$ is 0.6, 0.8 and 1. The suspected reason for these outliers is the collinearity in covariate matrix. Intuitively speaking, as the size of motivation increases, the average cost values decrease. The algorithm will update its selection strategy to choose to receive an intervention more likely. In this way, the collinearity of cost feature increases.

$$C_t = 10 - S_{t,1} + 0 \times S_{t,2} + 0 \times S_{t,3} - A_t \times (0.25 + 0.25S_{t,1} + 0.4S_{t,2}) + \xi_{t,4}$$

At each decision point, when $A_t = 1$, the cost feature is the vector $[1, S_{t,1}, S_{t,2}, S_{t,3}, S_{t,1}, S_{t,2}, S_{t,3}]$. When $A_t = 0$, the cost feature is $[1, S_{t,1}, S_{t,2}, S_{t,3}, 0, 0, 0]$. The condition number is the largest eigenvalue of the cost feature matrix. We want the square root of the condition number of our cost feature matrix to be close to 1. A large value of the square root of the condition number indicates high collinearity. Figure 4 provides the square root of condition number of cost feature matrix under different levels of physical motivation when T_{max} is 200. As $x_{coefficient}$ increases, the collinearity of the cost feature increases, and the corresponding IQR and MAD of policy

values increase as indicated by table 4 and5.

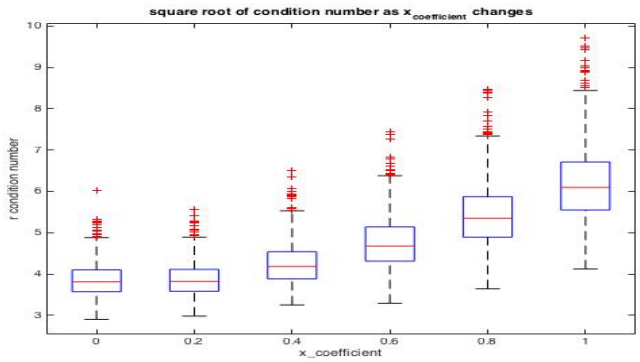


Figure 4: Square Root of Condition Numbers as $x_{coefficient}$ Increases

We suspect that the outliers are special cases. We repeated the simulation in model.b when $x_{coefficient}$ is 0.8 and 1 with a different seed. Figure 5 shows that there is no significant outlier in this simulation.

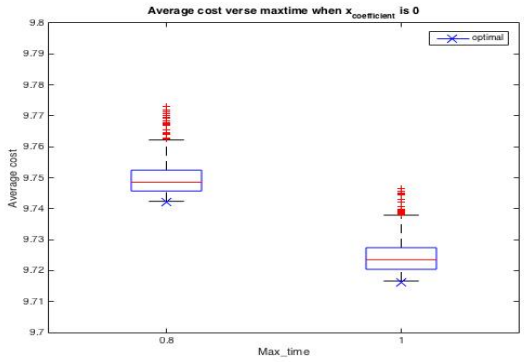


Figure 5: Average Cost With A Different Simulation Seed

5.2.5 Convergence of Policy Values and Policy Parameters

From the simulation results of model.a, it is observed that the performance of the contextual bandit algorithm can be improved by increasing the observation time T_{max} for each simulated person. In this section, we increase the observation time for different levels of motivation effects. We set $x_{coefficient} = 0$, 0.6, and 1, and observe the performance of the algorithm as the observation time T_{max} increases from 200 to 1000.

Figure 6 shows that for each value of $x_{coefficient}$, the average cost values get closer to the corresponding optimal value, and the variance of the average cost decreases when the observation time increases. This observation suggests that when a physical activity motivation effect exists, the algorithm can still improve its performance by increasing observation time for each person.

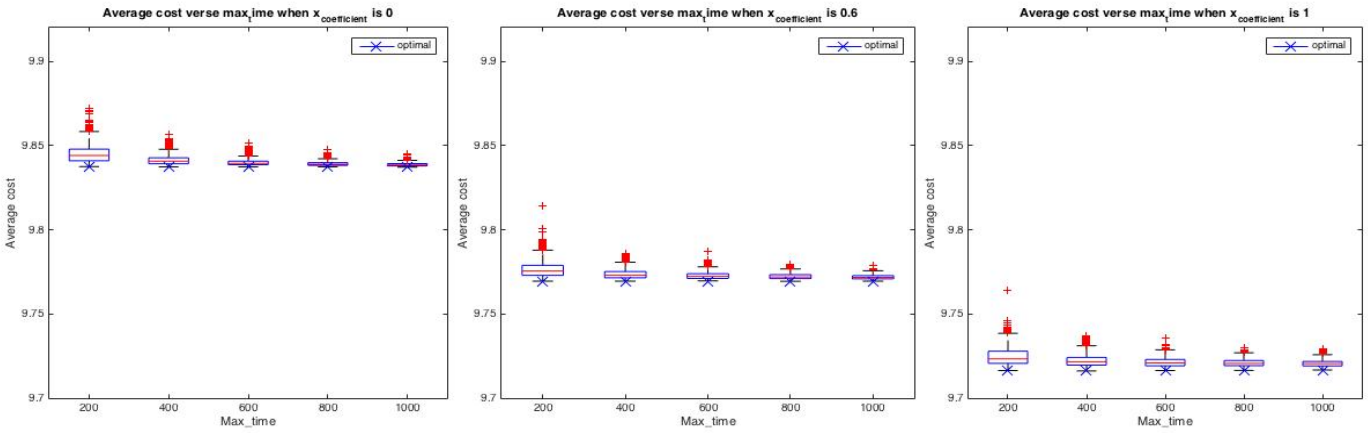


Figure 6: Average cost vs. T_{max}

The MAD of average cost values for $x_{coefficient} = 0$, 0.6, 1 also decreases as T_{max} increases. In addition, there is no significant increase in MAD when $x_{coefficient}$ increases from 0 to 1 under given T_{max}. The corresponding median absolute deviation values are provided in table 7.

Table 7: Median Absolute Deviation

$$MAD = median_i(|X_i - median_j(X_j)|)$$

$x_{coefficient}$	T_{max}				
	200	400	600	800	1000
0	0.0044	0.0022	0.0014	0.0011	0.0008
0.6	0.0038	0.0022	0.0017	0.0014	0.0012
1	0.0044	0.0028	0.0022	0.0018	0.0016

Since the performance of the algorithm can be improved by increasing the number of decision points, we expect to see that the policy parameters converge when T_{max} increases. The tables 8, 9, 10 below provide the bias and MSE of policy parameters when $x_{coefficient}$ is 0, 0.6 and 1 respectively. From the tables, we observe that for each $x_{coefficient}$, the bias of each $\hat{\theta}$ is stable when T_{max} increases, and the variance of each $\hat{\theta}$ decreases when T_{max} increases. This observation suggests that the policy parameters converge, but do not converge to the optimal policy.

Table 8: Bias and MSE of policy parameters when $x_{coefficient}$ is 0

$$Bias(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

T_{max}	θ_0	θ_1	θ_2	θ_3
200	-0.0011	-0.0067	-0.0029	-0.0017
400	-0.0009	-0.0052	-0.0022	-0.0004
600	-0.0027	-0.0045	-0.0017	0.0008
800	-0.0030	-0.0025	-0.0014	-0.0011
1000	-0.0017	-0.0023	-0.0017	-0.0008

$$MSE = Var(\hat{\theta}) + Bias(\hat{\theta})^2$$

T_{max}	θ_0	θ_1	θ_2	θ_3
200	0.0159	0.0109	0.0116	0.0108
400	0.0079	0.0054	0.0060	0.0053
600	0.0050	0.0037	0.0038	0.0036
800	0.0036	0.0026	0.0028	0.0028
1000	0.0028	0.0021	0.0021	0.0022

Table 9: Bias and MSE of policy parameters when $x_{coefficient}$ is 0.6

$Bias(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$

T_{max}	θ_0	θ_1	θ_2	θ_3
200	-0.0766	-0.0001	0.0057	-0.0005
400	-0.0742	-0.0005	0.0048	-0.0005
600	-0.0762	-0.0011	0.0051	0.0004
800	-0.0761	-0.0002	0.0051	-0.0010
1000	-0.0755	0.0000	0.0050	-0.0005

$MSE = Var(\hat{\theta}) + Bias(\hat{\theta})^2$

T_{max}	θ_0	θ_1	θ_2	θ_3
200	0.0147	0.0051	0.0051	0.0047
400	0.0097	0.0025	0.0026	0.0022
600	0.0085	0.0017	0.0017	0.0015
800	0.0078	0.0012	0.0013	0.0012
1000	0.0073	0.0009	0.0010	0.0009

Table 10: Bias and MSE of policy parameters when $x_{coefficient}$ is 1

$Bias(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$

T_{max}	θ_0	θ_1	θ_2	θ_3
200	-0.1217	0.0023	0.0070	-0.0020
400	-0.1200	0.0021	0.0059	-0.0012
600	-0.1220	0.0020	0.0058	0.0001
800	-0.1228	0.0028	0.0058	-0.0010
1000	-0.1223	0.0028	0.0058	-0.0006

$MSE = Var(\hat{\theta}) + Bias(\hat{\theta})^2$

T_{max}	θ_0	θ_1	θ_2	θ_3
200	0.0234	0.0039	0.0037	0.0034
400	0.0186	0.0020	0.0019	0.0017
600	0.0175	0.0013	0.0012	0.0011
800	0.0168	0.0009	0.0009	0.0009
1000	0.0163	0.0007	0.0008	0.0007

6 Discussion

6.1 Observation and Conclusion

In the hope of applying the contextual bandit algorithm to mobile health for providing optimal activity suggestions, this study explored the robustness of the contextual bandit algorithm

when deviation to its i.i.d assumption happens. Since a physical activity motivation effect is concerned as a major factor affecting the distribution of context states, this study explored the impact of a physical activity motivation effect on the performance of the algorithm through simulation and comparison. The results of our simulation study imply that the algorithm can still effectively choose the optimal policy. In addition, the impact of a physical activity motivation effect on decreasing the average sedentary time can be captured by the algorithm. Thus, based on the observations from simulation results, we claim that the physical activity motivation effect that buried in the contexts evolving process do not nullify the application of the contextual bandit algorithm.

6.2 Limitation

Physical activity motivation is not the only factor that violates the assumption of the contextual bandit algorithm. Even though the algorithm performs well in our generative model, in another generative model in Huitian Lei's work [1], where the motivation effect also impacts the cost generation, there is more bias in the policy parameters. Moreover, in real world mobile health problems, there are various kinds of ways that the distribution of contexts can be impacted by past interventions. For example, a disengagement to an intervention represented by S_3 in our model.b is also a major factor in the practical world affecting the context states distribution. When the disengagement is present, the average sedentary time is expected to increase. Further study may involve the disengagement factor in the cost generation model to observe the performance of the algorithm.

References

[1] Lei, Huitian, Ambuj Tewari, and Susan Murphy. "An Actor-Critic Contextual Bandit Algorithm for Personalized Interventions Using Mobile Devices." 2-4. Web.

[2] Li L, Chu W, Langford J, & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. Proceedings of the 19th International Conference on World Wide Web - WWW 10, 661-661.

[3] Bauer, Stephanie, et al. "Technology-enhanced maintenance of treatment gains in eating disorders: efficacy of an intervention delivered via text messaging." Journal of consulting and clinical psychology 80.4 (2012): 700.

[4] 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.

[5] Klasnja, Predrag, et al. "Microrandomized trials: An experimental design for developing just-in-time adaptive interventions." Health Psychology 34.S (2015): 1220.

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

[7] Nahum-Shani, Inbal, Eric B. Hekler, and Donna Spruijt-Metz. "Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework." Health Psychology 34.S (2015): 1209.

[8] 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.

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