Introduction and motivation

- Mobile health (mHealth): the practice of medicine and public health supported by mobile devices (wireless).
- Just in time adaptive intervention (JTIA): an intervention that are delivered in real-time, and are adapted to the immediate and changing needs of individuals as they go about their daily lives (Nahum-Shani et al 2014), by employing the real-time data collection and communication capabilities that modern mobile devices provide.
- JTIA utilizes a policy that takes patient information as input and outputs recommended interventions.
- We make a first attempt to bridge the methodological gap between the increasing popularity of JTIA as an improvement in real-world health care and the lack of methodological guidance in constructing data-based high-quality JTIA.
- We provide an online actor-critic algorithm to learn the optimal JTIA.

Framework: contextual bandits and parsimonious stochastic policy

Consider a series of pre-chosen decision points \{1, 2, ..., T\} where user-device interaction happens. Upper case letter to denote random variable, lower case letter to denote realization.
- Finite action space \(A\): all suitable interventions at the time (different types, dose, modalities of intervention, etc). Denote \(A_t\) the action at time \(t\).
- Context space \(S\): vector space of intervention relevant summary information (GPS location, self-reported measures, etc). Denote \(S_t\) the context vector at time \(t\).
- Linear expected reward: Given context \(s\) and action \(a\), reward is a linear function of a \(d\) dimensional feature vector \(f(s,a)\) with a unknown coefficient vector \(\mu\) plus a noise term \(\epsilon\) with standard deviation \(\sigma\).

\[
R = f(s,a)^T \mu + \epsilon
\] (1)

The choice of \(\theta_t\) 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 \(t+1\).
- Our focus of the policy is a class of parameterized stochastic policies \(\pi(\theta, s, a)\) where \(\theta \in \mathbb{R}^p\) is a \(p\) dimensional vector.

\[
\pi(\theta, s, a) = \frac{e^{f(s,a)^T \theta}}{\sum_{a'} e^{f(s,a')^T \theta}}
\] (2)

1. Contribution of each element in context is reflected by \(\theta\).
2. Creating intervention variety (exploration) preventing intervention burden/habituation/boredom.
3. Most likely \(p < d\).

Comparing to existing contextual bandit literature

1. We focus on low dimensional policy space. Scientists to choose a small set of contextual variables that potentially moderate the effect of interventions.
2. We encourage a minimal amount of exploration primarily to improve intervention adherence and enhance intervention effectiveness. Algorithmic benefits are secondary.
3. We quantify data uncertainty, thus the uncertainty of the learned optimal policy, by performing statistical inference on policy parameters, such as creating asymptotic confidence intervals.

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.

\[
V(\theta) = \sum_{s \in S} \pi(\theta, s, a) E[R(s,a)]
\]

Naturally, we define the optimal policy \(\pi_\star\), as the policy that maximizes the average reward, \(\theta_\star = \text{argmax}_\theta V(\theta)\). We argue for the need to penalize

- When \(E[R(s,a)] = s, A-a)\) is a (near) constant for all \(a\), maximizing \(V(\theta)\) in an ill-posed problem since solution is not unique.
- When \(E[R(s,a)] = s, A-a)\) is not a constant, it is easy to provide example where the maximizer of \(V(\theta)\) has one or more entries equal to \(\infty\). One or more infinitely large \(\theta\)'s render other \(\theta\)'s unidentifiable and the problem ill-posed (Mahanobhy 2012)

Our solution to circumvent ill-posed problems is to regularize the average reward by subtracting a \(\lambda\) penalty term.

- The regularized average reward is

\[
J(\theta) = V(\theta) - \lambda \| \theta \|^2
\] (3)

where \(\lambda\) is a hyperparameter that controls the amount of regularization. The optimal policy is \(\pi_\star(\theta, s, a)\) with \(\theta = \text{argmin}_\theta J(\theta)\).

- Maximizing \(J(\theta)\) becomes a well-posed problem and convergence to a pure stochastic policy is guaranteed in the absence of intervention effects.

- Guarantees a minimum exploration probability.

An actor-critic algorithm with linear expected reward

Algorithm 1: Actor critic algorithm for linear expected reward

1. Start from \(t = 0\).
2. At decision point \(t\), observe context \(s_t\).
3. Draw an action \(a_t\) according to probability distribution \(\pi_{B_t}(a_t|s_t)\).
4. Observe an immediate reward \(r_t\).
5. Critic update: \(B(t+1) = B(t) - \alpha (f(s_t,a_t)^T \theta_t - R_t) A(t)\).
6. Actor update: \(\theta_t = \text{argmax}_{\theta} \sum_t \sum_a \pi_{\theta}(s_t,a_t) f(s_t,a_t)^T \theta_t - R_t \|\theta\|^2\)

\[
\theta_t = \text{argmax}_{\theta} \sum_t \sum_a \pi_{\theta}(s_t,a_t) f(s_t,a_t)^T \theta_t - R_t \|\theta\|_2^2
\] (4)

Go to decision point \(t+1\).

Theory: consistency, rate of convergence and asymptotic distributions

Theorem 1. The critic's estimate \(\hat{\theta}\) converges to \(\theta^*\) in probability. The convergence rate is\(O(1/\sqrt{T})\), the optimal parameter convergent rate. Furthermore, \(\hat{\theta}(t) = A(t)^{-1} B(t)\) converges in distribution to multivariate normal with mean \(\theta^*\) and covariance matrix \((E[R(s,a)])/2(\theta^*)^2\).

Theorem 2. The critic's estimate \(\hat{\theta}\) converges to \(\theta^*\) in probability. The convergence rate is\(O(1/\sqrt{T})\). Furthermore, \(\hat{\theta} = A(t)^{-1} B(t)\) converges in distribution to multivariate normal with mean \(\theta^*\) and covariance matrix \(\sigma^2 E[R(s,a)](\theta^*)^2\).

Simulation: An example of reducing college students smoking

- Test the actor critic algorithm using an example in the context of reducing smoking among college students (Witkiewitz et al 2014).
- Relevant contextual information, \(s = [s_1, s_2, s_3, s_4]\): \(s_1 = \) smoking urge, \(s_2 = \) fixed mind, \(s_3 = \) indicator of low smoking mood, \(s_4 = \) indicator of user not busy.
- Three decision points per day. Action: \(a = 0\) provide informational intervention, \(a = 1\) provide behavioral intervention.
- Given context \(s\) and action \(a\), the negative reward, or the cost is the smoking rate \(C\). Smoking rate defined by the average cigarettes smoked per hour between two decision points.

\[
C = 6.54 + 1.16s_1 - 1.97s_2 + 2.08s_3 - 1.47s_4 - 2.87s_4 + 1.425s_1 + s_{16} + 1.060s_2 + 1.25s_4 + \epsilon
\]

where \(\epsilon\) is a truncated normal random variable.

- Consider the class of policies \(\pi_c(\theta, s, a) = a^c\theta_1 + (1-a^c)\theta_2\).

Acknowledgement

We acknowledge the support of University of Michigan Medical grant for Methodology for Developing Real-Time Mobile Phone Adaptive Interventions, NIH grants 5P50DA010075, and Grant US4EB020404 (NIDCR, NIAMS, NHLBI, NIBIB, NIAA).