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

Huitian Lei, Ambuj Tewari and Susan Murphy

Introduction and motivation

- Increasing technological sophistication and widespread use of smartphones and wearable devices provide opportunities for innovative health interventions, in particular, real-time personalized interventions.
- Just in time adaptive intervention(JITAI) are interventions that are delivered in real-time, and are adapted to address 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.
- A first attempt to bridge the methodological gap between the increasing popularity JITAIs receive from clinical and behavioral science community and the lack of methodological guidance in constructing data-based high quality JITAI.
- An online actor-critic algorithm is proposed to learn the optimal JITAI.

Problem formulation: contextual bandits and parametrized stochastic policies

Consider a sequence of decision points $\{1, 2, ..., t, ...\}$ when user-device interaction happens. Upper case letter to denote random variable, lower case letter to denote realization.

1. Finite action space $\mathcal{A}$: all suitable interventions at the time (different types, dose, modalities of intervention, etc). Denote A_t the action at time t.
2. Context space $\mathcal{S}$: vector space of intervention relevant summary information (GPS location, self-reported measures, etc). Denote S_t the context vector at time t.
3. Linear expected reward: Given context s and action a, reward is a linear function of a d dimensional reward feature vector $f(s, a)$ with an unknown coefficient vector μ^* plus a noise term ϵ with standard deviation σ.

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

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 policy in order to improve its choice of action at decision point $t+1$.

4. A class of parameterized stochastic policies $\pi_\theta(A = a|S = s)$ where $\theta \in \mathbb{R}^p$ is a p dimensional vector. When the action is binary

$$P(A = 1|S = s) = \pi_\theta(A = 1|S = s) = \frac{e^{g(s)^T\theta}}{1 + e^{g(s)^T\theta}} \tag{2}$$

- Interpretable policy: contribution of each context in the vector $g(s)$ is reflected by θ.

Regularized average reward

The average reward of a policy $\pi_\theta(A|S)$ is :

$$V^*(\theta) = \sum_{s\in\mathcal{S}} d(s) \sum_{a\in\mathcal{A}} f(s, a)^T \mu^* \pi_\theta(A = a|S = s)$$

Lemma 1. *(Deterministic optimal policy) Suppose the context space is discrete and finite, $\mathcal{S} = \{s_1, s_2, ..., s_K\}$. All s_j's are positive. The policy class is $\pi_\theta(A = 1|S = s) = \frac{e^{\theta_0+\theta_1 s}}{1+e^{\theta_0+\theta_1 s}}$. There exists a maximizer of $V^*(\theta)$ for which both θ_0 and θ_1 are infinite. In other words, one of the maximizers of $V^*(\theta)$ is a deterministic policy.*

Problems with deterministic policies:

- Increased burden/habituation/boredom by repeating the same intervention in the same context.
- When there is not enough exploration, the learned policy may be stuck at sub-optimal policy .

For References, Questions and comments please contact ehlei@umich.edu

- One infinitely large policy parameter renders all other parameters unidentifiable. The policy is no longer interpretable.

Introduce stochastic (chance) constraint: with high probability (probability taken with respect to the context) the probability of taking each action is bounded away from 0 :

$$P_S(p_0 \le \pi_\theta(A = 1|S)) \ge 1 - \alpha \tag{3}$$

By applying the Markov inequality, we reach a relaxed and smoother stochastic constraint that produces computational tractability:

$$\theta^T \mathbb{E}[g(S)^T g(S)]\theta \le (\log(\frac{p_0}{1-p_0}))^2 \alpha \tag{4}$$

Solving the constraint maximization problem (maximizing the average reward $V^*(\theta)$ subject to constraint (3)) by maximizing the Lagrangian function (also called as the regularized average reward function):

$$J^*_\lambda(\theta) = \sum_{s\in\mathcal{S}} d(s) \sum_{a\in\mathcal{A}} f(s, a)^T \mu^* \pi_\theta(A = a|S = s) - \lambda\theta^T \mathbb{E}[g(S)^T g(S)]\theta \tag{5}$$

An online actor-critic algorithm

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. $A(0) = 0_d$ is a $d \times 1$ column vector.
Actor initialization: θ_0 is the best treatment policy based on domain theory of historical data.
Start from $t = 0$.
while $t \le T_{max}$ **do**
  At decision point t, observe context S_t ;
  Draw an action a_t according to probability distribution $\pi_{\theta_{t-1}}(A|S_t)$;
  Observe an immediate reward R_t ;
  Critic update:
  $B(t) = B(t-1) + f(S_t, A_t)f(S_t, A_t)^T$, $A(t) = A(t-1) + f(S_t, A_t)R_t$, $\hat{\mu}_t = B(t)^{-1}A(t)$. ;
  Actor update:

$$\hat{\theta}_t = \underset{\theta}{\operatorname{argmax}} \frac{1}{t} \sum_{\tau=1}^{t} \sum_a f(S_\tau, a)^T \mu_t \pi_\theta(A = a|S_\tau) - \lambda\theta^T [\frac{1}{t} \sum_{\tau=1}^{t} g(S_i)^T g(S_i)]\theta \tag{6}$$

  Go to decision point $t + 1$.
end

Theory: consistency, rate of convergence and asymptotic distributions

Lemma 2. *Given a fixed λ, the population optimal policy θ^* lies in a compact set. In addition, the estimated optimal policy $\hat{\theta}_t$ lies in a compact set with probability going to 1.*

Theorem 1. *(Asymptotic properties of the critic) The critic's estimate $\hat{\mu}_t$ converges to μ^* in probability. The convergence rate is $O(1/\sqrt{t})$, the optimal parametric convergence rate. Furthermore, $\sqrt{t}(\hat{\mu}_t - \mu^*)$ converges in distribution to multivariate normal with mean 0_d and covariance matrix $[\mathbb{E}_{\theta^*}(f(s, a)f(s, a)^T)]^{-1}\sigma^2$, where $\mathbb{E}_\theta(f(s, a)f(s, a)^T) = \sum_s d(s) \sum_a f(s, a)f(s, a)^T \pi_\theta(a|s)$. The plug-in estimator of the asymptotic covariance is consistent.*

Theorem 2. *(Asymptotic properties of the actor) The actor's estimate $\hat{\theta}_t$ converges to θ^* in probability. The convergence rate is $O(1/\sqrt{t})$. Furthermore, $\sqrt{t}(\hat{\theta}_t - \theta^*)$ converges in distribution to multivariate normal with mean 0_p and covariance matrix $(J_{\theta\theta}(\mu^*, \theta^*)^{-1} V^* (J_{\theta\theta}(\mu^*, \theta^*))^{-1}$, where $V^* = \sigma^2 J_{\theta\mu}(\mu^*, \theta^*)\mathbb{E}_\theta(f(s, a)f(s, a)^T J_{\mu\theta}(\mu^*, \theta^*) + \sum_s d(s) j_\theta(\mu^*, \theta^*, s) j_\theta(\mu^*, \theta^*, s)^T$. The plug-in estimator of the asymptotic covariance is consistent. In the expression of asymptotic covariance matrix,*

$$j(\mu, \theta, S) = \sum_a f(S, a)^T \mu \pi_\theta(A = a|S) - \lambda\theta^T [g(S)g(S)^T]\theta$$

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

$J_{\theta\theta}$ and $J_{\theta\mu}$ are the second order partial derivatives of J. j_θ is the first order partial derivative of j.

Bootstrap Confidence Interval for the Policy

- A sample of reward noise is created by $\{\epsilon_t = R_t - f(S_t, A_t)^T \mu_T\}_{t=1}^T$
- A bootstrap sample for the estimated optimal policy θ^b_T and for the variance estimate $\hat{V}^b$ is generated according to algorithm 2
- We create bootstrap percentile-t confidence intervals for θ^*_i:

$$[\hat{\theta}_i - p_\alpha \frac{\hat{V}_i}{\sqrt{t}}, \hat{\theta}_i + p_\alpha \frac{\hat{V}_i}{\sqrt{t}}] \tag{7}$$

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

Start from $t = 0$. Initialize θ^b_0 to be the optimal policy based on domain theory
while $t < T$ **do**
  Context is s_t ;
  Bootstrap an action a^b_t according to probability distribution $\pi_{\theta_{t-1}}(s_t, a)$;
  Bootstrap the residuals to generate a bootstrapped reward $r^b_t = f(s_t, a^b_t)^T \mu_T + \epsilon^b_t$
  Critic update:
  $\mu^b_t = (\sum_{\tau=1}^T f(s_\tau, a^b_\tau)f(s_\tau, a^b_\tau)^T)^{-1}(\sum_{\tau=1}^T f(s_\tau, a^b_\tau)r^b_\tau)$;
  Actor update:

$$\theta_t = \underset{\theta}{\operatorname{argmax}} \frac{1}{t} \sum_{\tau=1}^{t} \sum_a f(s_\tau, a)^T \mu^b_t \pi_\theta(s_\tau, a) - \lambda\theta^T [\frac{1}{t} \sum_{\tau=1}^{t} g(s_\tau, 1)^T g(s_\tau, 1)]\theta$$

  Go to decision point $t + 1$
end
Plugin μ^b_T and θ^b_T to the formula in Theorem 2 to get a bootstrapped variance estimate $\hat{V}^b$.

Some Simulations

Generative model. The cost (negative reward) is generated according to

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

. We study the class of policies that include all contexts as potential tailoring variables: $\pi_\theta(A = 1|S = [S_1, S_2, S_3]) = \frac{e^{\theta 0+\theta_1 S_1+\theta_2 S_2+\theta_3 S_3}}{1+e^{\theta 0+\theta_1 S_1+\theta_2 S_2+\theta_3 S_3}}$. All simulation results shown below are based on the 1000 simulated datasets. All bootstrapped CI's, with a common nominal confidence level 0.95, are built based on 500 bootstrap samples.

We set $\tau = 0.4$. Contexts $\{[S_{t,1}, S_{t,2}, S_{t,3}]\}_{t=1}^T$ are iid multivariate normal with mean 0 and identity covariance matrix. With $\alpha = 0.1$, $p_0 = 0.2$, the optimal policy is $\theta^* = [0.65, 0.62, 0.98, 0]$. Table 2 shows the coverage rate of bootstrapped CI for sample sizes 200 and 500.

T(sample size)	θ_0	θ_1	θ_2	θ_3
200	0.941	0.912	0.915	0.939
500	0.943	0.940	0.954	0.941

Table 1: iid, identity covariance matrix, $\alpha = 0.1$, $p_0 = 0.2$, $\tau = 0.4$

We acknowledge the support of University of Michigan Mcubed grant for Methodology for Developing Real-Time Mobile Phone Adaptive Interventions, NIH grants P50DA010075, and Grant U54EB020404 (NIDCR, NIAMS, NHLBI, NIBIB, NIA).