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

Paper Authors: Huitian Lei, Ambuj Tewari, Susan Murphy
Presenter: Halley Young

Application domain: Just in Time Adaptive Interventions (JITAI)

- Behavioral “micro-interventions” via mobile phone app for behavioral health challenges such as smoking/alcohol cessation, increasing physical exercise, treating eating disorders, etc.
- Basic idea: App tracks both automatically gathered and user reported data, decides at a given time point whether to send out an encouragement to do a positive behavior or avoid a negative behavior
- Contextual bandit: We have a sequence of time steps to try using or not using interventions, only have information about the reward of the action we chose

Application domain advantages over certain other Contextual Bandit problems

- "Micro-interventions" - the only reward received from taking an action is the immediate reward (or lack thereof) - no direct effect of A_t on R_{t+n}, n >= 1
    - this is true in most contextual bandit problems, but worth repeating because it is surprising in this context but valid according to domain experts
- The set of features in the context is fairly small
    - in the case of physical activity suggestion, only take into account weather, activity level in past 30 minutes, and minimal self-reported data
    - compare to news articles dealing with thousands of different keyword searches

Application domain advantages over certain other Contextual Bandit problems

- The action space is small
  - choose to make it a 2-armed bandit problem - either do or don't make suggestion of behavior modification at time t

Application domain complications over certain other Contextual Bandit problems

- Need policy to be stochastic to prevent user habituation
- Need policy to be interpretable to please conductors of studies
  - compare to multilayer NNs used in similar contexts
- Feature data very autoregressive (esp. weather)

The Actor-Critic Formulation of the Contextual Bandits Problem

- A response to the need to make the policy interpretable
- 2 parts: an actor who chooses a policy, and a critic who evaluates the chosen policy
- Come up with a simple formula with only a 4-dimensional vector parameter to decide probability of choosing 0/1 (the “actor” parameter), but make the model to decide what that parameter should be (based on what reward it will produce, the “critic” parameter) more complicated

Model

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

where g(s) =
[1, weather goodness, current activity, current engagement with app]

Model

Problem - could be deterministic, which would cause habituation!
Add constraint:

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

“In at least (1-a)*100% of the possible contexts, there is at least a p0 and at most a (1 - p0) change of getting a 1 according to the policy.”

(But the real math is actually harder due to non-convexity of previous constraint)

Regularized average reward:

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

λ is fixed parameter for determining how much stochasticity to impose.

Then:

$$\theta_\lambda^* = \operatorname{argmax} J_\lambda^*(\theta)$$

The optimization in practice

Maximize the following:

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

Note: this is the form of the reward maximized by the actor, and assumes that we know $r(S_t, a)$ (which is the critic's job!)

The actor-critic algorithm

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 $f(s,a)^T \hat{\mu}_t$
>
> **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

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

The actor-critic algorithm

Note that the critic relies on f(s,a) = [], which is a significantly more informative feature vector than g(s) and includes features depending on both s (the context) and a (the action)

$$f(S_t, A_t) = [1, S_{t,1}, S_{t,2}, S_{t,3}, A_t, A_tS_{t,1}, A_tS_{t,2}, A_tS_{t,3}]$$

Optimality Guarantees: Additional Assumptions

1) The p-by-p matrix $\mathbb{E}(g(S)g(S)^T)$ is positive semidefinite
2) There exist constants which bound the possible features and reward at a time t (true in contexts like number of steps taken in 30 minutes!)
3) Linear model works with $R_t = f(s,a)^T\mu^* + \epsilon_t$ for some u* and normally distributed error term ε
   - note that this one will be tested empirically as well,, but is necessary for theoretical guarantees!

Optimality Guarantees: Critic

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.*

Optimality Guarantees: Actor

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:

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

Numerical Experiments

Simplest Experiment

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

Choose actual values for policy parameters, and model how long it takes and how well the actor in the actor-critic approximates this policy

Results:

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

Experiment 2: Autoregressive Contextual Features

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

$$\xi_{t,1} \sim N(0, 1-0.4^2),\ \xi_{t,2} \sim N(0, 1-0.4^2) \text{ and } \xi_{t,3} \sim N(0,1)$$

Results:

Convergence as T increases to 0

Bootstrap Confidence interval of θ3 is much lower (which is weird maybe because it’s the only non-autoregressive feature?)

Experiment 3: Actions Cause Increased Burden

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

Here C_t is cost (so trying to minimize rather than maximize reward)

τ represents the amount of burden associated with engagement

Results: As burden levels go up, MSE in optimal policy goes up dramatically (which authors claims is due to "overtreatment")

Experiment 4: Expected Cost is a nonlinear function of Cost feature f(s,a)

Model:

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

α controls how much nonlinearity (cost becomes quadratic in $S_{t,i}$)

MSE inflates, confidence intervals deteriorate as level of nonlinearity increases

Any Questions?